UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Técnicas de representação semântica de imagens por Bags of Complex Signatures - BoCS

Endi Daniel Coelho Silva

Dissertação de Mestrado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Endi Daniel Coelho Silva

Técnicas de representação semântica de imagens por Bags of Complex Signatures - BoCS

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Agma Juci Machado Traina

USP – São Carlos
Maio de 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S586t

Silva, Endi Daniel Coelho
Técnicas de representação semântica de imagens por Bags of Complex Signatures - BoCS / Endi Daniel Coelho Silva; orientadora Agma Juci Machado Traina. -- São Carlos, 2023.
83 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Imagem. 2. Banco de dados. 3. Redes complexas. 4. Semântica. I. Traina, Agma Juci Machado, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Endi Daniel Coelho Silva

Semantic Representation Techniques for Images using Bags of Complex Signatures - BoCS

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Agma Juci Machado Traina

USP – São Carlos
May 2023

AGRADECIMENTOS

Agradeço à minha orientadora Profa. Dra. Agma Juci Machado Traina.

Aos amigos da república, João Victor de Oliveira Novaes e Thiago de Jesus de Oliveira Durães.

Minha família, pelo apoio durante o tempo de mestrado.

Agradeço ao apoio financeiro da Coordenação de Aperfeiçoamento de Pessoa de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

SILVA, E. D. C. **Técnicas de representação semântica de imagens por Bags of Complex Signatures - BoCS**. 2023. 83 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Devido ao grande número de imagens geradas no nosso cotidiano que devem ser armazenadas e utilizadas posteriormente, surgiu a necessidade da criação de mecanismos eficientes para o gerenciamento de imagens em grandes bases de dados. Os sistemas de recuperação de imagens baseada em conteúdo são uma das principais ferramentas para realizar esse tipo de tarefa. No entanto, esse tipo de sistema muitas vezes não atende à expectativa do usuário, pois a forma como as imagens são representadas (características de baixo nível, como textura, forma e cor) não conseguem representar a semântica do pensamento humano. A representação de imagens utilizando palavras visuais, também conhecida por *Bag of Visual Words* (BoVW), é uma das mais famosas formas de tentar representar semanticamente o conteúdo de imagens. Portanto, o objetivo deste trabalho é desenvolver um método baseado no paradigma BoVW que consiga representar as imagens carregando informação semântica, possibilitando que essas representações possam ser utilizadas em sistemas de recuperação de imagens baseada em conteúdo e tragam resultados que sejam mais próximos às expectativas do usuário.

Palavras-chave: Imagens, Semântica, recuperação de imagens baseada em conteúdo, Bag of Visual Words.

ABSTRACT

SILVA, E. D. C. **Semantic Representation Techniques for Images using Bags of Complex Signatures - BoCS**. 2023. 83 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Due to the large number of images generated in our daily lives that must be stored and used later, there is a need to create efficient mechanisms for managing images in large databases. Content-based image retrieval systems are one of the main tools to perform this type of task. However, this type of system often does not meet the user's expectations, because of the way images are represented (low-level characteristics, such as texture, shape and color) cannot represent the semantics of the humans' thought. The representation of images using visual words also known as Bag of Visual Words (BoVW), is one of the most famous ways of trying to represent the content of images semantically. Therefore, the objective of this work is to develop a method based on the BoVW paradigm that is able to represent images carrying semantic information, enabling these representations to be used in content-based image retrieval systems and bring results that are more consistent with what the user want.

Keywords: Images, Semantic, Content-based image retrieval, Bag of Visual Words.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BoCS	*Bag of Complex Signatures*
BOSS	Bag-Of-SuperPixel Signatures
BoVW	*Bag of Visual Words*
C-BoVW	Cluster-Based Bag-of-visual-words
CBIR	Content Based Image Retrieval
CBIR	Recuperação de Imagens Baseada em Conteúdo
GLCM	Gray Level Co-occurrence Matrix
GSA	*Global Spatial Arrangement*
LBP	*Local Binary Pattern*
MAP	*Mean Average Precision*
S-BoVW	*Signature-Based Bag of Visual Words*
SDLC	*Sorted Dominant Local Color*
SDLCT	*Sorted Dominant Local Color and Texture*
SDLT	*Sorted Dominant Local Texture*
SGBD	Sistemas Gerenciadores de Banco de Dados
SIFT	“Scale-invariant feature transform”
SLIC	*Simple Linear Iterative Clustering*
SURF	“Speeded Up Robust Features”
TBIR	Text Based Image Retrieval

LISTA DE SÍMBOLOS

D — Domínio dos dados complexos

U — Domínio dos vetores de características

F — Extrator de Características

I — Imagem

H — Histograma

μ — média

σ — variância

α — assimetria

V — Conjunto de vértices do grafo

E — conjunto de arestas do grafo

G — Grafo

r — raio

L — Maior valor de intensidade encontrado em uma imagem

T — Conjunto de limiares

C — Dicionário de palavras visuais

B — Conjunto de blocos

ε — Limiar de frequência

Q — quadrante

Z — Conjunto de histogramas ordenados

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

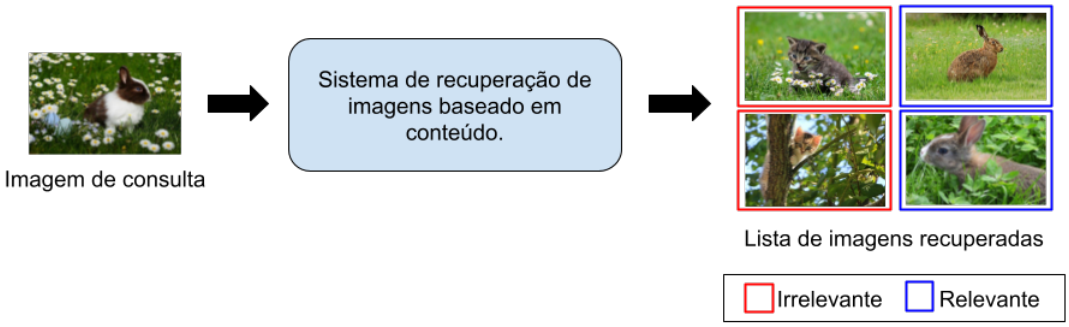
As imagens são elementos fundamentais em diversas áreas da ciência e da sociedade. No jornalismo, elas são usadas para prender a atenção dos leitores com cenas impactantes, enquanto na educação, são empregadas para apresentar informações por meio de gráficos. Já na saúde, as imagens são imprescindíveis para armazenar resultados de exames, como os de Raio-X e tomografias. Devido à sua importância, os avanços nas tecnologias de obtenção e armazenamento têm gerado um aumento significativo no número de imagens geradas a cada ano. Assim, é necessário o uso de sistemas eficientes para manipulação, armazenamento e recuperação de imagens em bancos de dados.

Para suprir a necessidade de recuperação de imagens relevantes em grandes bases de dados, surgiram sistemas de recuperação de imagens baseados em texto (TBIR - *Text Based Image Retrieval*) e baseados em conteúdo (CBIR - *Content-Based Image Retrieval*) . Na abordagem TBIR, o conteúdo visual das imagens é descrito por anotações feitas por seres humanos, que são utilizadas para indexar e recuperar as imagens. No entanto, essa abordagem tem limitações, já que a descrição do conteúdo visual pode ser subjetiva e variar de acordo com a interpretação e o conhecimento do anotador. Por outro lado, na abordagem CBIR, as imagens são representadas por vetores de características extraídas automaticamente a partir das informações visuais, que são usados para indexar, buscar e comparar as imagens. Isso permite que os sistemas CBIR realizem a recuperação de imagens de forma mais precisa e eficiente, sem depender da subjetividade humana. Nessa abordagem, as imagens são representadas por vetores de características utilizados para indexar, buscar e comparar as imagens. Normalmente esses vetores são obtidos a partir de uma análise das características visuais da imagem, que pode ser feita de forma global, observando a imagem como um todo, ou de forma local, identificando na imagem regiões de interesse e obtendo as características em cada uma das regiões (ALKHAWLANI; ELMOGY; EL-BAKRY, 2015).

1.1 Motivação e Justificava

Quando os usuários utilizam sistemas CBIR, é esperado que as imagens retornadas sejam relevantes para a consulta realizada. Neste trabalho, consideramos como imagens relevantes aquelas que pertencem à mesma classe da imagem de consulta, conforme ilustrado pela Figura 1. Nesse exemplo, as imagens destacadas em vermelho pertencem à classe dos gatos, enquanto as destacadas em azul são de coelhos. Portanto, se um usuário inserir uma imagem de um coelho no sistema, ele espera que sejam retornadas imagens da mesma classe (ou seja, outras imagens de coelhos).

Figura 1 – Exemplo de busca por similaridade em um sistema de recuperação de imagens baseado em conteúdo.



Fonte: Elaborada pelo autor.

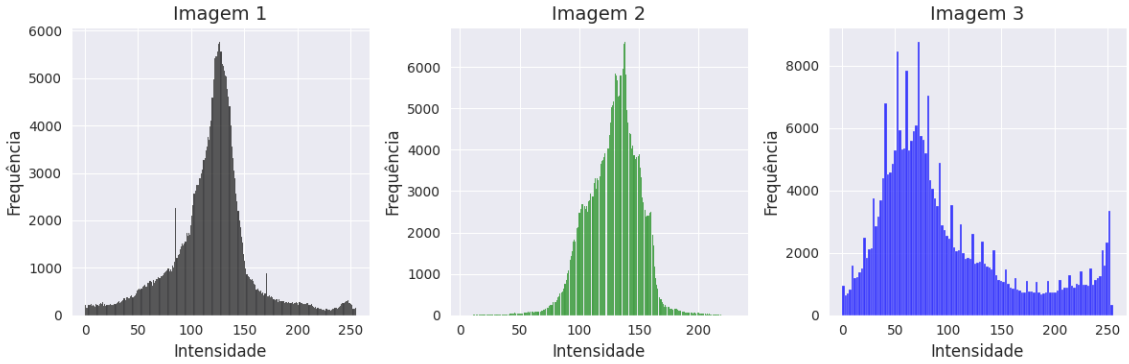
Os sistemas computacionais geralmente utilizam características como cor, textura e forma extraídas dos pixels das imagens para gerar suas representações e realizar comparações para determinar as imagens mais similares (ALEMU *et al.*, 2009). Uma das formas de se gerar representações de imagens seria analisar sua distribuição de cores, ou tipo de textura, no entanto esse tipo de abordagem não é o suficiente para diferenciar as imagens de forma eficiente(ALEMU *et al.*, 2009). A Figura 2 ilustra um exemplo onde imagens com distribuições de cores parecidas pertencem a classes diferentes, podemos ver que o gato da coluna 1, possui uma distribuição de seus tons de cinza visualmente mais similar ao coelho na coluna 2 do que o gato na coluna 3. Vale ressaltar que avaliar somente o formato da distribuição não é o suficiente. Uma das maneiras mais utilizadas na literatura para esse tipo de comparação seria o cálculo da distância entre suas representações.

A ideia por trás desse exemplo, é trazer uma intuição do porque utilizar somente a distribuição de cores (ou intensidades) para buscar imagens similares pode não ser uma boa ideia. Uma alternativa para resolver esse problema seria analisar as regiões das imagens de forma local, avaliando regiões de interesse por toda a imagem e gerando diferentes representações para cada região. Dessa forma é possível capturar diferenças específicas que existem em regiões das imagens, no entanto isso gera um número muito grande de vetores de características o que dificulta o uso desse tipo de representações em sistemas CBIR (PEDROSA, 2015).

Figura 2 – Exemplo de como observar apenas a distribuição de cores de uma imagem pode levar a resultados que não correspondem ao esperado pelo usuário.

(a) Imagens que serão analisadas a distribuição de cores

(b) Distribuição dos tons de cinza das imagens



Fonte: Elaborada pelo autor.

O método *Bag of Visual Words* (BoVW) destacou-se na literatura por sua capacidade de sumarizar as características locais das imagens em um único vetor de características. Em sua abordagem tradicional, as regiões de interesse são descritas localmente e, a partir disso, os vetores que representam cada região passam por um processo de agrupamento, resultando em um dicionário de assinaturas visuais, onde normalmente o centro de cada grupo corresponde a uma assinatura visual. A imagem é descrita contabilizando o número de assinaturas visuais pertencentes a cada grupo, gerando um histograma de palavras visuais que a representa. Devido à sua capacidade de analisar as características locais e gerar uma representação simples que pode ser usada em sistemas CBIR, as abordagens baseadas no BoVW serão estudadas neste trabalho.

1.2 Objetivos

Este trabalho tem como objetivo principal desenvolver e avaliar uma abordagem para geração de assinaturas de imagens que possam ser utilizadas em sistemas CBIR (recuperação de imagens por conteúdo) visando recuperar imagens similares às imagens fornecidas como consulta.

1.2.1 Objetivos Específicos

- Avaliar diferentes extratores de características para identificação das características visuais das imagens.
- Investigar abordagens baseadas no método BoVW (Bag-of-Visual-Words) para geração de assinaturas de imagens.
- Explorar extensões do método BoVW que possam melhorar sua representação de características visuais.
- Investigar e avaliar funções de distância que possam ser utilizadas em conjunto com as assinaturas geradas para realizar a recuperação de imagens por conteúdo.
- Avaliar o desempenho das assinaturas geradas em ambientes de CBIR, considerando diferentes conjuntos de dados e métricas de avaliação.

1.3 Organização do Trabalho

O texto desta monografia está organizado da seguinte forma: no Capítulo 2 são apresentados os conceitos básicos necessários ao desenvolvimento da pesquisa, no Capítulo 3 os trabalhos relacionados à proposta de pesquisa são detalhados, enquanto no Capítulo 4 é apresentada a proposta de pesquisa para o Mestrado e por fim no Capítulo 5 apresentamos as conclusões e futuras linhas de pesquisa.

CAPÍTULO

2

CONCEITOS

Neste capítulo, serão apresentados os conceitos básicos que são essenciais para o entendimento da pesquisa de mestrado. Esses conceitos incluem sistemas de recuperação de imagens baseados em conteúdo (CBIR), Extratores de características, Redes complexas, Consultas por similaridade e os diferentes paradigmas do método *Bag of Visual Words*. Entender esses conceitos é fundamental para a compreensão das técnicas e metodologias utilizadas na pesquisa em questão, bem como para o desenvolvimento de sistemas eficientes de recuperação de imagens. Portanto, este capítulo é fundamental para o desenvolvimento de uma base sólida de conhecimento e para o sucesso da pesquisa de mestrado.

2.1 Recuperação de Imagens por Conteúdo - CBIR

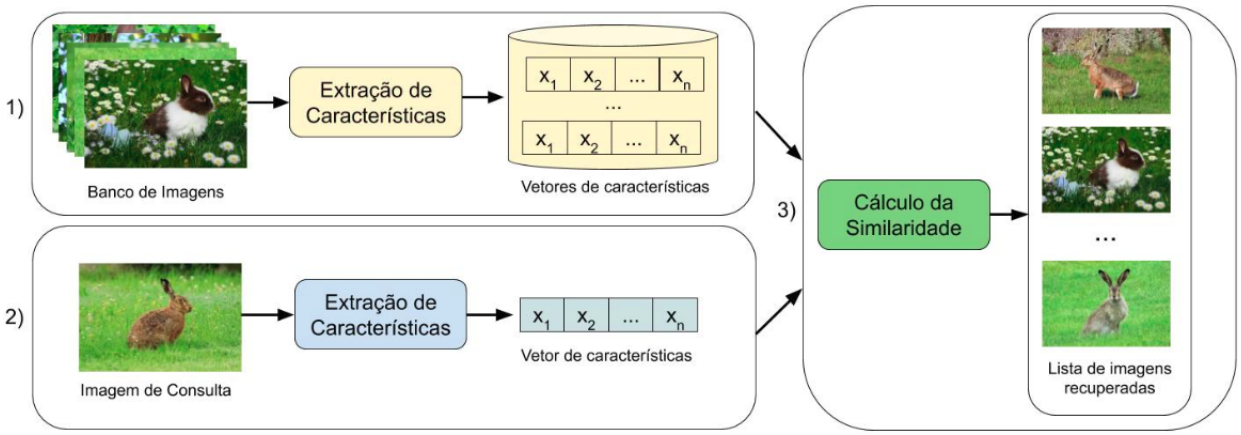
Devido ao grande número de imagens geradas no nosso cotidiano, surgiu a necessidade da criação de mecanismos eficientes para o gerenciamento de imagens em grandes bases de dados. A Recuperação de Imagens Baseada em Conteúdo (CBIR) se destaca na literatura como ferramenta para gerenciar imagens em grandes bases de dados sem depender de anotações manuais ou tags. Na medicina, ela pode ser usada para identificar imagens médicas semelhantes, facilitando o diagnóstico e o tratamento de doenças(VISHRAJ; GUPTA; SINGH, 2022). No comércio eletrônico, essa técnica pode ser utilizada para facilitar a busca e a compra de produtos on-line. Os usuários podem procurar produtos com base em suas características visuais, como cor e forma, em vez de depender apenas de descrições de texto (ALSMADI, 2020)

O processo de recuperação de imagens por conteúdo pode ser dividido em três etapas, a primeira é a extração de características, na qual as características visuais das imagens são identificadas e extraídas. Diversos descritores podem ser utilizados, tais como cor, textura e forma, dependendo das necessidades da aplicação. A segunda etapa é a indexação das imagens, na qual elas são organizadas de forma a permitir uma busca eficiente. Por fim, a terceira etapa é a busca por similaridade, na qual as imagens são comparadas com base nas características

extraídas e na estrutura de indexação criada na etapa anterior. Cada uma dessas etapas é crucial para o sucesso do processo de recuperação de imagens por conteúdo. A escolha cuidadosa dos descritores de características, técnicas de indexação e busca por similaridade pode aumentar significativamente a precisão e a eficiência da recuperação de imagens por conteúdo. (JIANG *et al.*, 2021)

A Figura 3 ilustra o funcionamento básico de um sistema CBIR. 1) No primeiro passo cada imagem da base de dados é representada por um vetor de características obtido ao se aplicar um extrator de características, dessa forma, sempre que for realizada uma consulta no sistema, os vetores de características das imagens já estarão calculados. 2) Ao receber uma imagem de consulta, um extrator de características é utilizado para obter o vetor que descreve as características visuais da imagem, segundo um critério determinado. 3) O vetor da imagem de consulta é comparado com todos os vetores das imagens pertencentes à base de dados utilizando uma função de distância/dissimilaridade e em seguida o conjunto das imagens mais similares à imagem de consulta é retornado para o usuário.

Figura 3 – Passo a passo de um sistema de recuperação de imagens por conteúdo.



Fonte: Elaborada pelo autor.

2.1.1 Extração de Características

A maioria dos Sistemas Gerenciadores de Banco de Dados (SGBD) não oferecem um suporte efetivo para operações de consulta, indexação e busca com dados complexos tais como imagens (SINGH; SINGH, 2020). Para que isso seja possível é necessário que as imagens sejam representadas por meio de um vetor que represente suas características visuais numericamente. Esse vetor é obtido ao se aplicar um extrator na imagem que pode utilizar desde recursos básicos, como cores e texturas, até recursos mais complexos, como a forma (DUBEY, 2022). A Figura 4 ilustra esse processo, nela temos uma imagem em que um extrator é aplicado e calcula valores numéricos que representaram as características visuais da imagem, tais como cor, forma ou textura que serão armazenados em um ou mais vetores de n dimensões, por fim a Definição 1 descreve formalmente como funciona um extrator de características.

Definição 1. Considere D como o domínio dos dados complexos, e U como o domínio dos vetores de características, um extrator de características é uma função $F : D \rightarrow U$ que transforma um dado $d_i \in D$ em $u_i \in U$. (CAZZOLATO, 2019)

Figura 4 – Funcionamento de um extrator de características.

Fonte: Elaborada pelo autor.

A escolha do extrator utilizado para representar as imagens varia conforme seu objetivo, em geral, a escolha do método depende da aplicação e das propriedades das imagens a serem processadas. Além disso, é importante escolher características que sejam discriminativas e robustas o suficiente para permitir a comparação entre imagens de forma eficaz (DUBEY, 2022). A seguir serão detalhados alguns dos mais famosos extratores utilizados na literatura.

2.1.1.1 Extratores de Cor

A informação de cor é uma das mais utilizadas para descrever visualmente uma imagem, entre as técnicas mais conhecidas para extrair esse tipo de informação está o histograma (SWAIN; BALLARD, 1991), obtido ao mapear as cores de uma imagem I em um espaço discreto contendo n cores. Formalmente um histograma de cores pode ser definido como um vetor $H = \{h_{c_1}, h_{c_2}, h_{c_3}, ..., h_{c_n}\}$ de forma que h_{c_i} indica quantos pixeis possuem a cor c_i na imagem. Essa é uma técnica simples e eficaz para representar as características de cor de uma imagem. No entanto, não leva em conta as informações de textura e forma da imagem, o que pode limitar sua precisão na recuperação de imagens por conteúdo.

Como alternativa ao histograma, foi proposto por (STRICKER; ORENGO, 1995) um extrator baseado em momentos de cores, ele utiliza os momentos estatísticos das distribuições de cores para descrever a informação cromática da imagem. Para que isso seja possível, considere N como o número de pixeis, i o canal de cores analisado e j o j-ésimo pixel da imagem, dessa forma $p_{i,j}$ representa a valor presente no pixel j e canal i, a partir disso, é possível calcular o primeiro momento(média) definido pela Equação 2.1 que indica o valor médio de cor na imagem, para um determinado canal i, além disso podemos calcular o segundo momento (variância) Equação 2.2 que nos diz como a cor varia na imagem e o terceiro momento (assimetria) Equação 2.3 nos diz o quão assimétrica é a distribuição de cores da imagem. Observe, que com essa abordagem, é

possível representar uma imagem por um vetor com três dimensões, uma para cada momento calculado sobre a imagem.

Os momentos de cor são uma técnica simples e eficaz para representar as características de cor de uma imagem. Eles têm sido amplamente utilizados em aplicações de recuperação de imagens por conteúdo, especialmente em combinação com outras técnicas de processamento de imagem, como o histograma de cor. No entanto, assim como o histograma de cor, os momentos de cor não levam em conta as informações de textura e forma da imagem, o que pode limitar sua precisão na recuperação de imagens por conteúdo.

$$\mu_i = \frac{1}{N}\sum_{j=1}^{n} p_{i,j} \tag{2.1}$$

$$\sigma_i = \sqrt[2]{\frac{1}{N}\sum_{j=1}^{n} (p_{i,j} - \mu_i)^2} \tag{2.2}$$

$$\alpha_i = \sqrt[3]{\frac{1}{N}\sum_{j=1}^{n} (p_{i,j} - \mu_i)^3} \tag{2.3}$$

2.1.1.2 Extratores de Textura

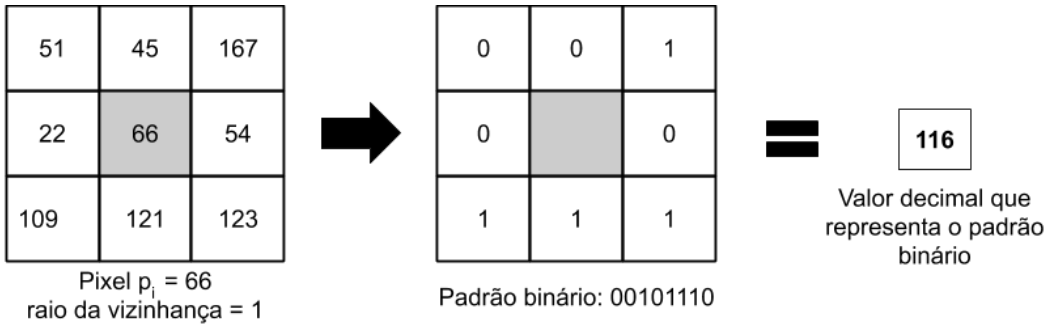
As características de textura são importantes em tarefas de reconhecimento de padrões e recuperação de imagens, e dependendo do tipo de imagens podem ser a principal forma de sua representação. Embora não exista uma definição formal, a textura intuitivamente é descrita por medidas que quantificam suas propriedades de suavidade, rugosidade e regularidade (GONZALEZ; WOODS, 2001). Uma das maneiras mais conhecidas para representar a textura em uma imagem é por meio da extração de medidas estatísticas da Gray Level Co-occurrence Matrix (GLCM). Essa matriz estuda a distribuição de níveis de cinza em uma imagem por meio da relação espacial entre os pares de pixeis. A matriz é construída observando a frequência relativa $p(i, j|d, \theta)$ entre os pixeis, isto é, o número de ocorrências na imagem em que pixeis de valor de intensidade i estão separados de pixeis de intensidade j à uma distância d na direção θ. A partir disso é possível extrair características para representar a textura da imagem. Haralick, Dinstein e Shanmugam (1973) propuseram um conjunto de 14 medidas para este propósito, no entanto, devido a alta correlação entre essas medidas Conners e Harlow (1980) mostraram que apenas 5 dessas medidas (*energy, entropy, correlation, local homogeneity, contrast*) eram necessárias.

Tamura, Mori e Yamawaki (1978) propuseram um conjunto de seis características de textura correlacionadas com a percepção visual humana (*Coarseness, Contrast, Directionality, Line-likeness, Regularity, Roughness*). Segundo Zhao, Xu e Hong (2009) essas seis características se destacam em aplicações nas quais o sistema de visão humano é utilizando como critério de performance.

O *Local Binary Pattern* (LBP) é um dos mais famosos extratores de características de textura encontrados na literatura, foi originalmente desenvolvido por Ojala, Pietikäinen e Harwood (1996) e representa cada pixel da imagem com um padrão binário. A ideia básica do LBP é considerar um pixel central em uma imagem e compará-lo com seus vizinhos em uma janela circular de determinado raio r. Para cada vizinho, é atribuído um valor binário 1 se o seu valor de intensidade é maior ou igual ao valor de intensidade do pixel central, ou 0 caso contrário. O padrão binário resultante é convertido em um número decimal que representa o padrão de textura local naquela região da imagem. Este processo é repetido para cada pixel na imagem, gerando uma matriz de números LBP. Uma vez que a matriz LBP é gerada, é possível extrair uma série de estatísticas descritivas dessa matriz, tais como histogramas, médias, variâncias, entropias e outros. Essas estatísticas podem ser usadas para classificar as texturas. A Figura 5 ilustra o processo de caĺculo do LBP. Formalmente o LBP de um determinado pixel p_i é calculado como indicado na Equação 2.4 no qual P indica o número de pixeis da vizinhança, e o valor de x é dado observando as seguintes condições: $x = 1$ se $p_i - p_j \geqslant 0$, caso contrário $x = 0$.

$$LBP_{p_i} = \sum_{j=0}^{P} x * 2^j \tag{2.4}$$

Figura 5 – Cálculo do LBP para um pixel da imagem.



Fonte: Elaborada pelo autor.

2.1.2 Medidas de Similaridade

Para medir a similaridade entre duas imagens, normalmente são utilizadas funções de distância, que mensuram o quão diferentes são duas imagens, ou seja, quanto maior o valor resultante da comparação entre duas imagens, menos similares elas são. Os vetores de características representam as propriedades visuais de uma imagem e, dependendo da aplicação, podem ser definidos de diversas maneiras. Por isso, diferentes tipos de vetores de características podem necessitar de medidas de similaridade distintas (PEDROSA, 2015). Selecionar qual medida de similaridade utilizar ao comparar duas imagens pode afetar significativamente a precisão de um sistema CBIR. Na literatura, funções de distância têm encontrado um bom nível de sucesso. No entanto, encontrar uma função adequada e robusta para medir a similaridade

entre duas imagens ainda é um desafio (ALZU'BI; AMIRA; RAMZAN, 2015). A seguir, serão apresentadas algumas funções de distância que podem ser utilizadas como medidas de similaridade.

Um tipo de função de distância amplamente utilizado em espaços vetoriais é o da família *Minkoswki*. Dados dois vetores de características em um espaço n-dimensional, a = $\{a_1, a_2, a_3, ..., a_n\}$ e b = $\{b_1, b_2, b_3, ..., b_n\}$, a família de distância *Minkoswki* pode ser definida de diferentes maneiras, variando apenas o parâmetro p, como definido pela Equação 2.5. A variação do parâmetro p nos leva a distâncias conhecidas na literatura: se $p = 1$, temos a distância *City-block*; se $p = 2$, temos a distância Euclidiana, e se p = ∞, temos a distância Chebychev (AVALHAIS, 2012).

$$distanciaMinkowski(a,b) = \sqrt[p]{\sum_{i=1}^{n} |a_i - b_i|^p} \tag{2.5}$$

A distância cosseno é calculada utilizando o produto interno dos dois vetores dividido pelo produto dos seus comprimentos (norma). A Equação 2.6 mostra como é feito este cálculo (SILVA, 2014).

$$distanciaCosseno(a,b) = 1 - \frac{a.b}{|a|.|b|} \tag{2.6}$$

A distância Canberra é calculada observando a diferença absoluta entre os atributos dos vetores e dividindo essa diferença pela soma absoluta de seus atributos, como pode ser visto na Equação 2.7 (SILVA, 2014).

$$distanciaCanberra(a,b) = \sum_{i=1}^{n} \frac{|a_i - b_i|}{|a_i| + |b_i|} \tag{2.7}$$

A distância Bray-Curtis é calculada de forma similar à Camberra como descrito na Equação 2.8 (SILVA, 2014).

$$distanciaBray - Curtis(a,b) = \sum_{i=1}^{n} \frac{|a_i - b_i|}{a_i + b_i} \tag{2.8}$$

Além das funções citadas anteriormente podemos citar a distância *Bhattacharyya* (KAILATH, 1967) que mede o quanto duas distribuições de probabilidade estão sobrepostas. Nesse contexto ela pode ser definida pela Equação 2.9.

$$distanciaBhattacharyya(a,b) = -\ln\left(\sum\left(\sqrt{a_i \cdot b_i}\right)\right) \tag{2.9}$$

A distância de Jaccard é calculada com base na sobreposição entre os conjuntos de características dos vetores. Ela mede a proporção de características compartilhadas em relação ao total de características únicas presentes em ambos os vetores (ZAHROTUN, 2016). Quanto

maior a sobreposição, maior será a similaridade entre as imagens. Para calcular a distância de Jaccard entre os vetores de características a e b, podemos utilizar a Equação 2.10. Onde $|a \cap b|$ representa o número de características compartilhadas pelos vetores(interseção) e $|a \cup b|$ representa o número total de características únicas presentes nos vetores (união).

$$distanciaJaccard(a,b) = 1 - \frac{|a \cap b|}{|a \cup b|} \tag{2.10}$$

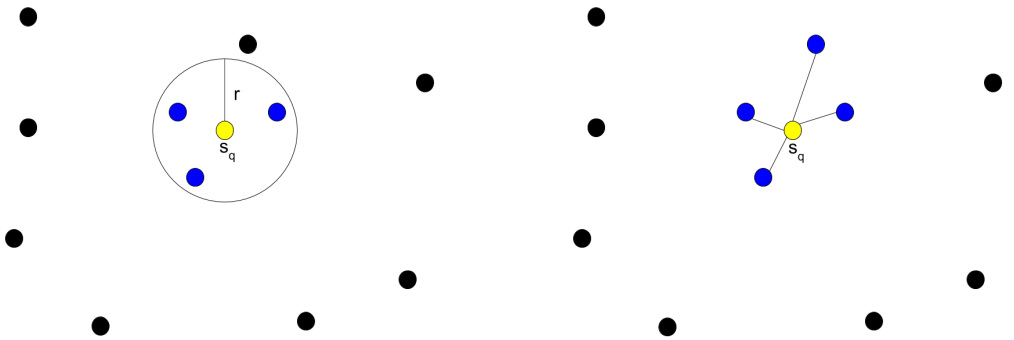
2.1.3 Consultas por similaridade

O objetivo de uma consulta por similaridade é recuperar elementos de uma base de dados que sejam similares a um dado elemento de consulta, na literatura os dois métodos mais comuns para se realizar este tipo de consulta são: a consulta por abrangência (*Range Query*) (ver Definição 2) e a consulta aos k-vizinhos mais próximos (*k-Nearest Neighbor - kNN*) (ver Definição 3).

Definição 2. *Consulta por abrangência*: dado um conjunto de dados $S = \{s_1, s_2, ..., s_n\}$, um elemento de consulta s_q, a consulta por abrangência irá retornar todos elementos de S que estiverem a no máximo uma distância r em relação a s_q. A Figura 6a ilustra o funcionamento deste tipo de consulta.

Definição 3. *Consulta aos k-vizinhos mais próximos*: dado um conjunto de dados $S = \{s_1, s_2, ..., s_n\}$ e um elemento de consulta s_q, a consulta aos k-vizinhos mais próximos retorna os k elementos mais similares a s_q em S. A Figura 6b ilustra o funcionamento deste tipo de consulta.

Figura 6 – Exemplo de consultas por similaridade sobre o elemento central s_q. Os objetos em azul indicam os elementos que resultam da consulta.



(a) Resultado de uma consulta por abrangência de raio r

(b) Resultado de uma consulta aos k-vizinhos mais próximos com $k = 4$

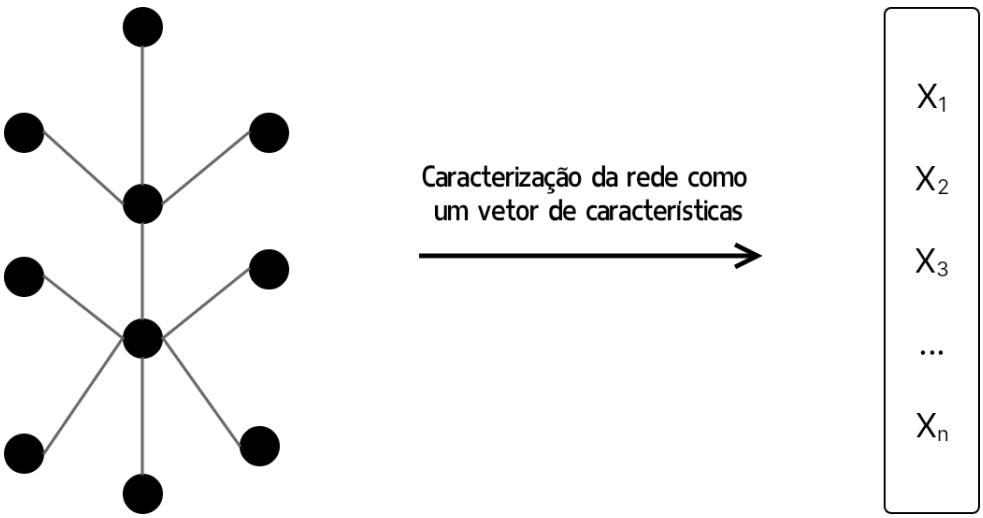
Fonte: Adaptada de Souza (2019, Página 45).

2.2 Redes Complexas

As redes complexas são sistemas compostos por um grande número de elementos conhecidos como vértices que são interconectados por arestas. Essas redes podem ser encontradas em diversas áreas, como biologia, ciência da computação, física, sociologia e economia. Elas apresentam diversas características, o que as torna um objeto de estudo interessante em diversas áreas de pesquisa. Uma das principais razões por trás de sua popularidade é a flexibilidade e generalidade para representar virtualmente qualquer estrutura natural como listas, árvores e imagens (COSTA *et al.*, 2007).

As pesquisas em redes complexas geralmente envolvem a representação da estrutura de interesse como uma rede, seguida de uma análise das características da representação obtida. Para redes pequenas, a visualização das imagens das redes é uma excelente maneira de entender sua estrutura, no entanto, à medida que essas redes crescem, essa abordagem se torna inviável. Portanto, estudos recentes utilizam métodos estatísticos, tais como comprimentos de caminho, grau médio de seus vértices, coeficiente de agrupamento médio, diâmetro e distribuições de grau, para analisar as características da rede (NEWMAN, 2003). Essa análise pode gerar uma representação em formato de um vetor de características, como ilustrado na Figura 7.

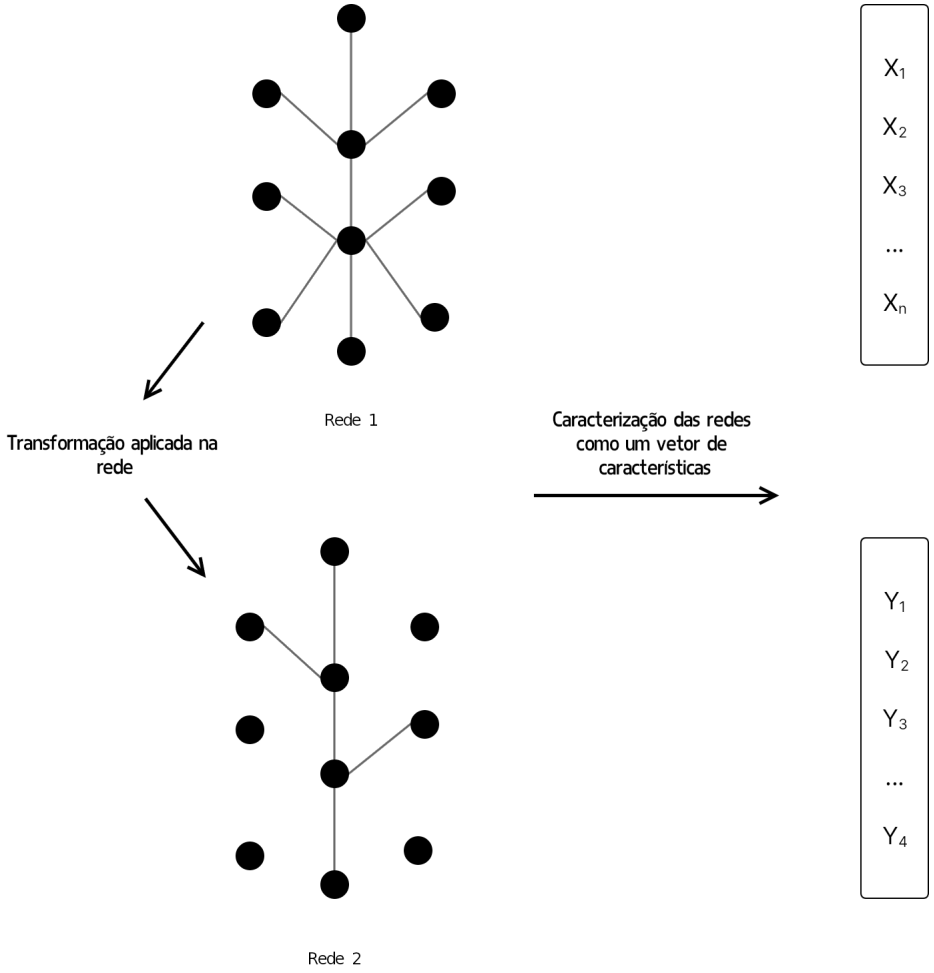
Figura 7 – Representação de uma rede complexa como um vetor de características.



Fonte: Adaptada de Costa *et al.* (2007, Página 5).

Uma estratégia que pode ser usada para obter informações adicionais sobre a estrutura de redes complexas envolve a aplicação de uma transformação na rede original e a obtenção das medidas da rede resultante, como ilustrado na Figura 8. Nesta figura, uma transformação é aplicada sobre a rede original deletando arestas com base em uma regra pré definida, para obter uma estrutura transformada a partir da qual novas medidas são extraídas (COSTA *et al.*, 2007).

Figura 8 – Transformação de uma rede 1 em uma rede 2 de forma que medidas estatísticas possam ser aplicadas para gerar representações com seus respectivos vetores de características.



Fonte: Adaptada de Costa *et al.* (2007, Página 5).

2.2.1 Medidas de redes complexas

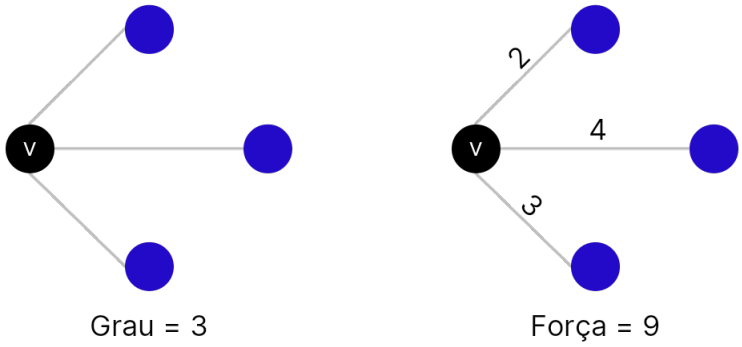
Como mencionado anteriormente, existem diversas medidas que podem ser aplicadas em redes complexas para gerar suas representações como vetores de características. Essas medidas estão relacionadas à conectividade dos vértices, distância, agrupamento, entre outras. A seguir, serão detalhadas algumas das medidas mais utilizadas para análise de redes complexas, divididas em duas categorias: locais e globais. As medidas locais são utilizadas para analisar a conectividade de cada vértice individualmente na rede, enquanto que as globais são utilizadas para caracterizar a rede como um todo.

2.2.1.1 Medidas locais

As medidas locais são úteis para entender como um vértice específico está conectado com outros em sua vizinhança, como ele é afetado pela remoção de outras arestas e como ele influencia a rede como um todo. O grau é uma das mais medidas mais simples quando vamos analisar os vértices de uma rede, pois ele representa o número de arestas conectadas ao vértice

em questão (COSTA *et al.*, 2007). Formalmente o grau k de um vértice v_i, pode ser calculado pela Equação 2.11, de forma que e_{v_i,v_j} representa uma aresta que conecta v_i à v_j. Para redes onde existe peso nas arestas a definição anterior pode ser utilizada para calcular a força s_i, no entanto ao invés de contar quantas arestas estão conectadas à v_i somamos o peso w_{v_i,v_j} das arestas, como podemos ver pela Equação 2.12. A Figura 9 ilustra um exemplo com o cálculo da força e o grau de um determinado vértice v.

Figura 9 – O Grau à esquerda definido pelo número de arestas conectadas ao vértice v, enquanto a força é calculada somando o peso das arestas conectadas à v



Fonte: Elaborada pelo autor.

$$k_i = \sum_j e_{v_i,v_j} \tag{2.11}$$

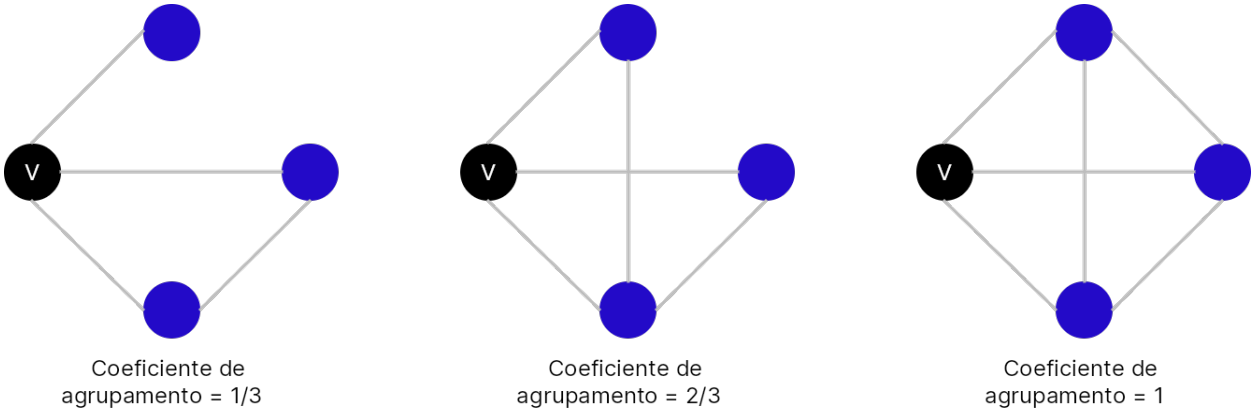
$$s_i = \sum_j w_{v_i,v_j} \tag{2.12}$$

Uma outra propriedade interessante para compreender a estrutura e o comportamento dos vértices da rede, são os coeficientes de agrupamento, eles nos dão a possibilidade de avaliar grupos dentro da rede. Um alto coeficiente de agrupamento de um vértice indica que seus vizinhos estão fortemente conectados entre si, o que pode indicar a presença de um grupo bem definido dentro da rede. Por outro lado, um baixo coeficiente de agrupamento de um vértice indica que seus vizinhos têm poucas conexões entre si, o que pode indicar que esse vértice está situado em uma região menos coesa com poucas conexões. Basicamente o coeficiente de agrupamento de um vértice v é definido pela razão entre o número de arestas que existem entre seus vizinhos e o número total de arestas possíveis. Seja k_v o grau de um determinado vértice v e l_v é o número de arestas existentes entre os vizinhos de v podemos definir seu coeficiente de agrupamento pela Equação 2.13 (WATTS; STROGATZ, 1998).

$$c_v = \frac{l_v}{k_v(k_v - 1)} \tag{2.13}$$

A centralidade de intermediação (ou *betweenness centrality*) é uma medida de centralidade em redes que mede a importância relativa de um vértice ou uma aresta como um

Figura 10 – Exemplos do cálculo do coeficiente de agrupamento, na primeira rede, a esquerda o coeficiente igual a 1/3 pois entre os vizinhos de v apenas existe 1 aresta de 3 possíveis, na rede do meio existem 2 arestas de 3 possíveis já na rede mais a direita o coeficiente é igual à um pois existem 3 arestas de 3 possíveis



Fonte: Adaptada de Arruda (2019, página 52).

intermediário no fluxo de informações entre outros vértices na rede (FREEMAN, 1978). Em outras palavras, a centralidade de intermediação indica o grau em que um vértice ou uma aresta é encontrado em caminhos mais curtos entre pares de vértices na rede, assim como definido pela Equação 2.14.

$$C_b(v) = \sum_{s \neq v \neq t} \frac{\sigma(s,t|v)}{\sigma(s,t)} \tag{2.14}$$

Nessa equação, $C_b(v)$ é a centralidade de intermediação do vértice v, $\sigma(s,t|v)$ é o número de caminhos mais curtos entre os vértices s e t que passam por v, e $\sigma(s,t)$ é o número total de caminhos mais curtos entre os vértices s e t. O somatório é realizado sobre todos os pares distintos de vértices s e t que não incluem o vértice v.

A centralidade de vértices é amplamente utilizada em diversas áreas, inclusive para análises de textura em imagens. Nesse contexto, cada pixel da imagem pode ser considerado um vértice do grafo e as arestas são definidas com base na diferença de intensidade entre os vértices. A centralidade pode ser usada para identificar regiões de com variações de intensidade de cor ou luminância muito acentuadas. Esses vértices com alta centralidade podem indicar regiões de interesse na imagem, como bordas ou áreas com texturas mais complexas, que geralmente apresentam variações de intensidade mais acentuadas.

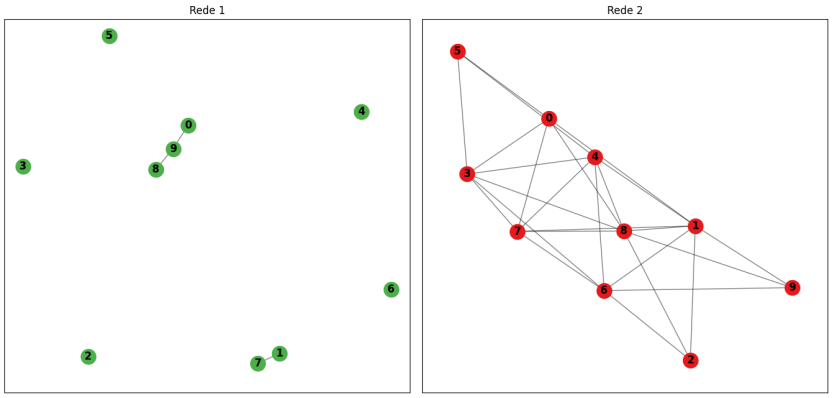
2.2.1.2 Medidas Globais

As medidas globais são fundamentais para análise de redes complexas, pois fornecem informações sobre sua estrutura e propriedades como um todo. Essas medidas possibilitam a comparação entre diferentes redes e a identificação de padrões gerais que podem ser relevantes

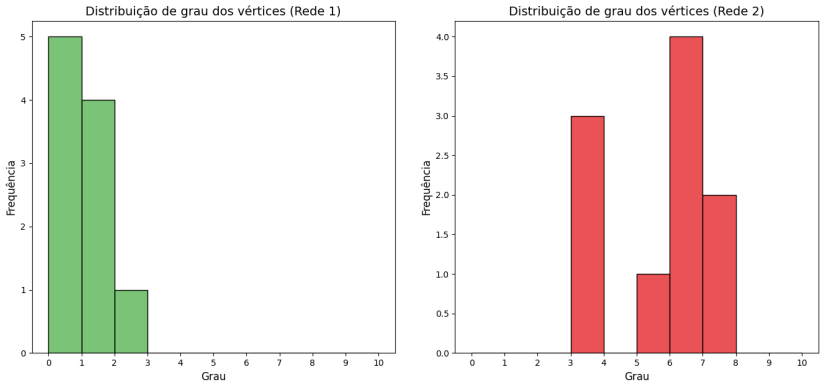
para entender o comportamento do sistema representado pela rede. Dentre as medidas globais mais simples, podemos destacar a distribuição de graus da rede, que mostra a fração de vértices com grau k, ou seja, a probabilidade de escolher um vértice de forma aleatória com grau k. Essa informação é útil para compreender a estrutura geral da rede, indicando se a rede é densamente conectada (alta frequência de vértices com muitas conexões) ou esparsamente conectada (alta frequência de vértices com poucas conexões), como ilustrado pela Figura 11. Além disso, a distribuição de força dos vértices pode ser utilizada para caracterizar a rede de forma similar (COSTA *et al.*, 2007).

Figura 11 – Visualização da distribuição de grau da rede

(a) A rede mais esquerda tem poucas arestas e portanto vértices com um grau menor, a rede a direia tem um conjunto maior de arestas e portanto vértices com graus maiores.



(b) A distribuição mais a esquerda mostra que a rede tem muitos vértices com baixo grau enquanto a distribuição mais a direita mostra que os vértices da rede possuem no geral um grau elevado.

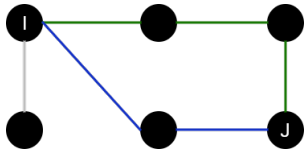


Fonte: Elaborada pelo autor.

Além da distribuição de graus, outra forma importante de analisar a estrutura de uma rede é medir o caminho mínimo médio entre seus vértices. Essa medida fornece informações sobre a eficiência da comunicação na rede, ou seja, o quão fácil é para os vértices se comunicarem entre si. Quanto menor o caminho mínimo médio, mais eficiente é a comunicação na rede. Em redes não ponderadas, o comprimento de um caminho é definido como o número de arestas necessárias para conectar dois vértices. Já em redes ponderadas, o comprimento de um caminho é a soma

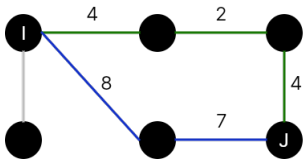
dos pesos das arestas no caminho(NEWMAN, 2003). A Figura 12 ilustra um exemplo em que há dois caminhos possíveis para ir do vértice i ao vértice j, sendo que o caminho em azul é o caminho mínimo d_{ij}. Já na Figura 13, que representa uma rede ponderada, o caminho mínimo é aquele que tem a menor soma de pesos de arestas, sendo o caminho em verde o caminho mínimo nesse caso.

Figura 12 – Caminho mínimo entre dois vértices em uma rede sem peso nas arestas.



Fonte: Elaborada pelo autor.

Figura 13 – Caminho mínimo entre dois vértices com peso nas arestas.



Fonte: Elaborada pelo autor.

A partir disso podemos definir o caminho mínimo médio entre os vértices da rede como a soma das distâncias mínimas entre os pares e dividida pelo numero total de pares de vértices como pode ser visto pela Equação 2.15 (COSTA *et al.*, 2007).

$$l = \frac{1}{N(N-1)} \sum_{i \neq j} d_{i_j} \tag{2.15}$$

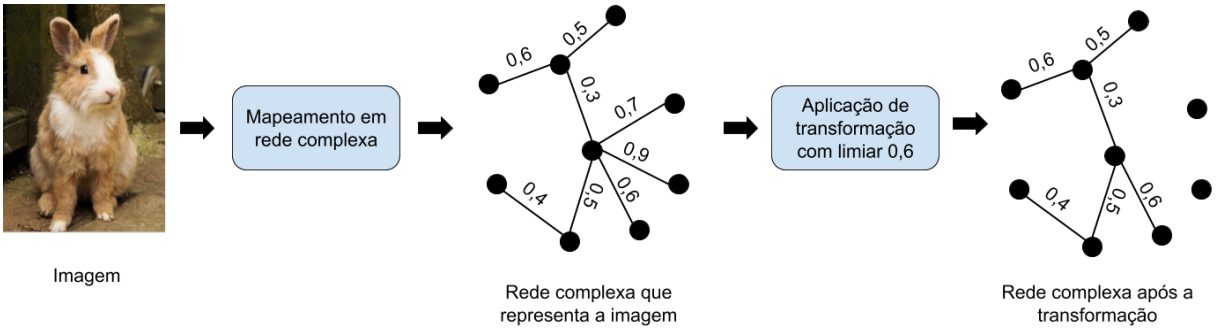
2.2.2 Modelagem de imagens como redes complexas

Como o objetivo principal deste trabalho é gerar representações de imagens para uso em sistemas CBIR, é importante explorar em detalhes como a modelagem de imagens como redes complexas é realizada e como ocorre o processo de extração de características visuais. Em artigos como Scabini, Gonçalves e Castro (2015) e Lima *et al.* (2019), esse processo é utilizado para extrair características de texturas. Como observado nesses trabalhos, a extração de características visuais da imagem ocorre em duas etapas. Na primeira etapa, a imagem é modelada como um grafo G. Na segunda etapa, medidas são extraídas de G para representar a imagem. Assim, uma imagem I pode ser representada por meio de um grafo $G = V, E$. Cada pixel $p_i \in I$ é representado como um vértice $v_i \in V$, e as arestas $e_{v_i,v_j} \in E$ indicam as conexões entre os pares de pixels que estão a uma distância de até r.

Após modelar a imagem como um grafo G, é possível extrair diversas medidas para representar cada pixel (vértice) localmente ou a rede como um todo. É importante destacar que

esse conjunto de medidas pode ser retirado da rede original ou de redes derivadas, obtidas a partir da aplicação de uma transformação. Um dos tipos de transformações mais comuns é a aplicação de um conjunto de limiares $T = t_1, t_2, ..., t_n$. Ao aplicar um limiar t_i na rede G, a rede derivada G_i manterá apenas as arestas que possuem peso menor ou igual a t_i. A Figura 14 ilustra esse processo.

Figura 14 – Aplicação de uma transformação em uma rede complexa, gerando uma rede derivada.



Fonte: Elaborada pelo autor.

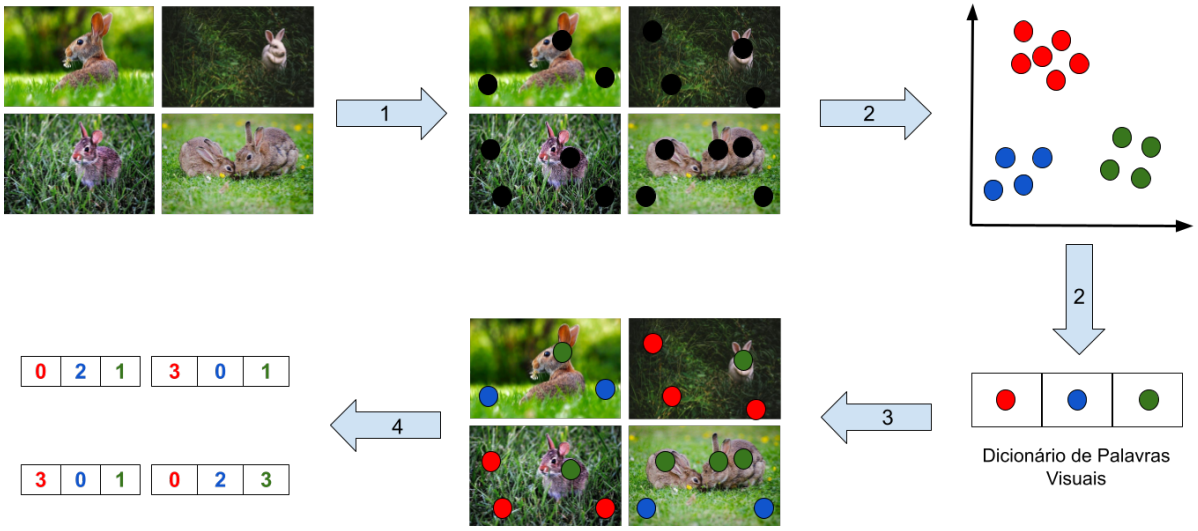
Em resumo, a modelagem de imagens como redes complexas e a extração de características visuais a partir dessas redes são técnicas que têm se mostrado promissoras na área de CBIR. Ao representar as imagens como grafos, é possível extrair medidas que capturam informações locais e globais da imagem. Além disso, as redes derivadas, obtidas a partir da aplicação de transformações, podem fornecer informações adicionais que complementam a representação original.

2.3 Discussão sobre a abordagem *Cluster-Based Bag-of-visual-words*

O Cluster-Based Bag-of-visual-words (C-BoVW) é uma abordagem inspirada no *Bag of Words*, técnica da área de recuperação textual que representa documentos por meio de um histograma de palavras. No C-BoVW o histograma é formado por palavras visuais, obtidas mediante o agrupamento das características locais de regiões de interesse. Essa representação permite aplicar técnicas de recuperação textual em imagens, o que facilita a busca de imagens em grandes coleções de dados (SANTOS, 2016).

A representação de imagens por meio de palavras visuais foi proposta em um dos primeiros trabalhos de detecção de objetos em vídeos por Sivic e Zisserman (2003). Desde então, essa abordagem tem sido amplamente utilizada em diversas aplicações. Por exemplo, Shamna, Govindan e Nazeer (2019) usam o C-BoVW para recuperar imagens médicas em bases de dados, enquanto Sun e Kise (2011) aplicam essa técnica para identificar plágio em mangás. A seguir será detalhada o funcionamento dessa técnica.

Figura 15 – Passo a passo bag of visual words



Fonte: Elaborada pelo autor.

Dado um conjunto de *m* imagens, a Figura 15 fornece uma visão de como funciona o C-BoVW em quatro passos: No passo 1) a detecção de regiões de interesse pode ser feita por meio de diferentes técnicas, como por exemplo a segmentação de imagens, detecção de bordas ou detecção de pontos de interesse. Já no passo 2), é comum utilizar descritores de características, como por exemplo o SIFT, SURF, para descrever cada região de interesse. No passo 3), a associação de cada região de interesse ao centroide mais próximo pode ser feita por meio de algoritmos de clusterização, como o k-means. Por fim, no passo 4), é possível montar um histograma que represente cada imagem pela frequência de palavras visuais.

É importante ressaltar que a escolha de técnicas e parâmetros em cada um dos passos pode influenciar na qualidade final da representação das imagens. Portanto, é importante avaliar a performance do C-BoVW em cada aplicação específica e realizar ajustes na metodologia conforme necessário.

2.3.1 Detecção das regiões de interesse

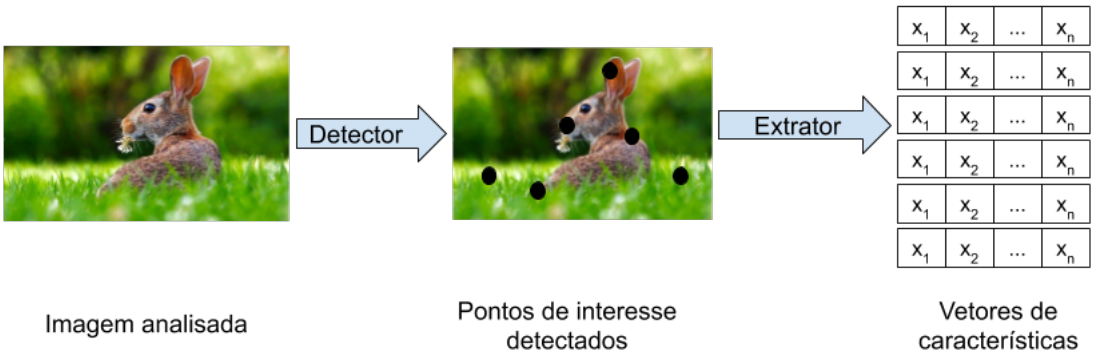
A detecção de regiões de interesse é um passo crucial no método C-BoVW para a representação de imagens. Ela permite extrair características visuais relevantes, utilizadas na criação do dicionário de palavras visuais. Nesta seção, serão apresentados os principais métodos para a detecção de regiões de interesse. Comumente, detectores de pontos de interesse são utilizados para identificar as áreas importantes de uma imagem. Esses detectores fornecem um número específico de localizações invariantes às transformações geométricas, ruídos e diferenças de iluminação, possibilitando que os mesmos pontos sejam identificados em imagens diferentes (MUKHERJEE; WU; WANG, 2015).

Existem diversas técnicas populares para realizar essa tarefa. Dentre elas, destacam-se

o método Harris-Affine (MIKOLAJCZYK; SCHMID, 2004) e as Diferenças de Gaussianas (LOWE, 1999), ambos baseados na detecção de características invariantes à rotação e escala. Além dessas técnicas mais avançadas, há a possibilidade de realizar a detecção de pontos de interesse de forma mais simples, como através do *Dense Sampling*, que detecta pontos igualmente espaçados na imagem, ou por meio de uma detecção aleatória, chamada de *Random Sampling* (PEDROSA, 2015). A escolha da técnica adequada depende das características das imagens e da aplicação em questão.

Após detectar os pontos de interesse é necessário utilizar algum extrator de características que represente suas informações visuais, como visto na Figura 16. Esses extratores geralmente trabalham com a vizinhança dos pontos de interesse. A escolha do extrator depende da aplicação e da base de imagens com que se está trabalhando.

Figura 16 – Detecção e extração de características dos pontos de interesse



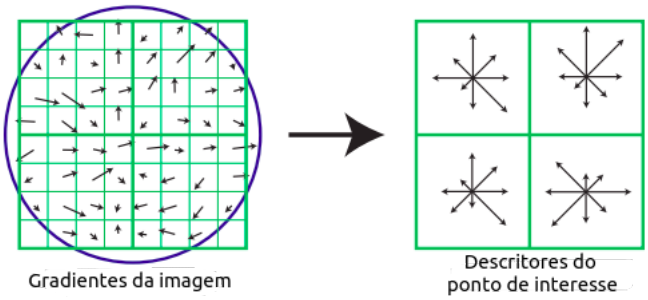
Fonte: Elaborada pelo autor.

Vale ressaltar que na literatura dois métodos se destacam por realizar o processo de detecção e representação dos pontos de interesse, o "Speeded Up Robust Features" (SURF) (BAY; TUYTELAARS; GOOL, 2006) e "Scale-invariant feature transform" (SIFT) (LOWE, 2004) sendo ambos amplamente utilizados em diversas aplicações.

O método SIFT foi baseado em um modelo de visão biológica, no qual neurônios no córtex visual primário respondem a um gradiente em uma orientação e frequência espacial específicas. O método consiste na utilização do Diferenças de Gaussianas (DoG) (LOWE, 1999) para detectar pontos de interesse na imagem, após isso, as características locais de cada ponto são descritas utilizando a magnitude e orientação do gradiente de seus pixeis vizinhos. Como ilustrado pela Figura 17, o primeiro passo é calcular a magnitude e orientação em cada pixel dentro de uma janela ao redor do ponto de interesse. A magnitude é ponderada por uma função gaussiana(indicada pelo círculo na imagem) de forma que quanto mais distante do centro menor o peso. Um histograma de orientações com 8 posições é criado para cada sub-região, de forma que cada posição do histograma corresponda à soma das magnitudes do gradiente próximas aquela direção na sub-região. Vale ressaltar que a Figura 17 mostra esse processo para um descritor 2x2

calculado de uma janela de 8x8. No entanto Lowe (2004) recomenda um descritor 4x4 utilizando uma janela 16x16 onde cada sub-região é representada por um histograma com 8 posições, o que resulta em um vetor de 128 elementos para representar o ponto de interesse.

Figura 17 – Descrição de um ponto de interesse utilizando o SIFT.



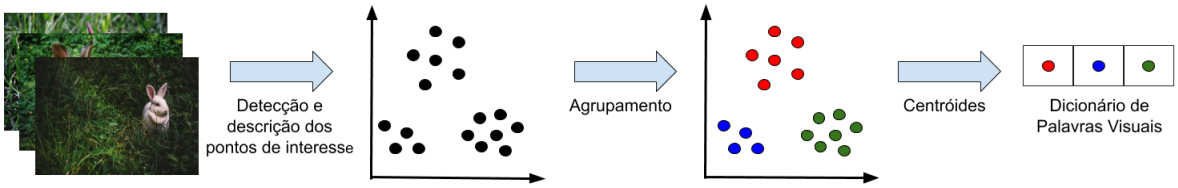
Fonte: Lowe (2004, Página 15).

O método SURF tem propriedades semelhantes ao SIFT, mas com uma complexidade reduzida. Entre as suas principais diferenças é possível citar o uso do *'Fast-Hessian'* baseado no *Hessian-Laplace* para detectar os pontos de interesse, o uso da transformada de *Haar* para descobrir a orientação dos pixeis vizinhos e o tamanho do vetor representativo final com 64 elementos enquanto o SIFT utiliza 128 elementos.

2.3.2 Dicionário de palavras visuais

Nesta etapa, o objetivo é construir um dicionário que possa representar as características e padrões presentes nas imagens de forma compacta. Para isso, após detectar e extrair as características dos pontos de interesse, é comum utilizar algoritmos de agrupamento, como o k-means, para se obter um número k de grupos. Cada centroide dos grupos será utilizado para representar uma palavra visual do dicionário. Dessa forma, a imagem pode ser representada pela frequência de ocorrência de cada palavra visual no dicionário. A Figura 18 ilustra esse processo.

Figura 18 – Passo a passo para a criação do dicionário de palavras visuais, primeiro ocorre a detecção dos pontos de interesse em seguida eles são agrupados e o centroide de cada grupo representa uma palavra visual no dicionário



Fonte: Elaborada pelo autor.

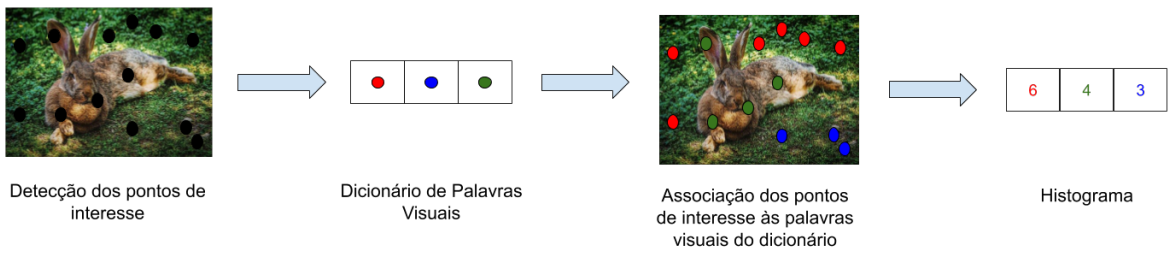
Portanto, seja $P = \{p_1, p_2, ..., p_n\}$ o conjunto dos pontos de interesse, A um método de agrupamento, o dicionário de palavras visuais pode ser definido como o conjunto $C = \{c_1, c_2, ...c_k\}$ dos centroides obtidos ao se aplicar A em P. Dessa forma, o dicionário de palavras

visuais é uma representação compacta das características das imagens, permitindo uma abstração de informações que torna possível a comparação entre imagens de maneira mais eficiente. O valor de k frequentemente é definido empiricamente, mas deve ser definido com cuidado, pois de acordo com Jiang, Ngo e Yang (2007), um dicionário pequeno pode ter pouco poder discriminativo, fazendo com que dois pontos de interesse distintos possam ser associados a uma mesma palavra visual, mesmo que não sejam semelhantes entre si, enquanto um dicionário grande tem pouco poder de generalização, é menos tolerante a ruídos e pode gerar processamento desnecessário. Por isso, é importante avaliar diferentes valores de k e escolher aquele que equilibra a discriminação e generalização do dicionário para a aplicação em questão.

2.3.3 Representação da imagem

Nesta etapa é feita a associação dos pontos de interesse a uma ou mais palavras visuais do dicionário C. O objetivo dessa associação é gerar um histograma H para cada imagem da base de dados utilizando palavras visuais. A abordagem mais comum para a criação do histograma é conhecida como *hard assignment*, nela a associação ocorre observando a proximidade entre os pontos de interesse e as palavras visuais do dicionário C. Dessa forma, cada ponto de interesse detectado na imagem é associado à palavra visual mais próxima no dicionário, a frequência das palavras visuais presentes na imagem formam o histograma que será interpretado como o vetor de características que representa a imagem. A Figura 19 ilustra esse processo.

Figura 19 – Contagem da ocorrências das palavras visuais em uma Imagem



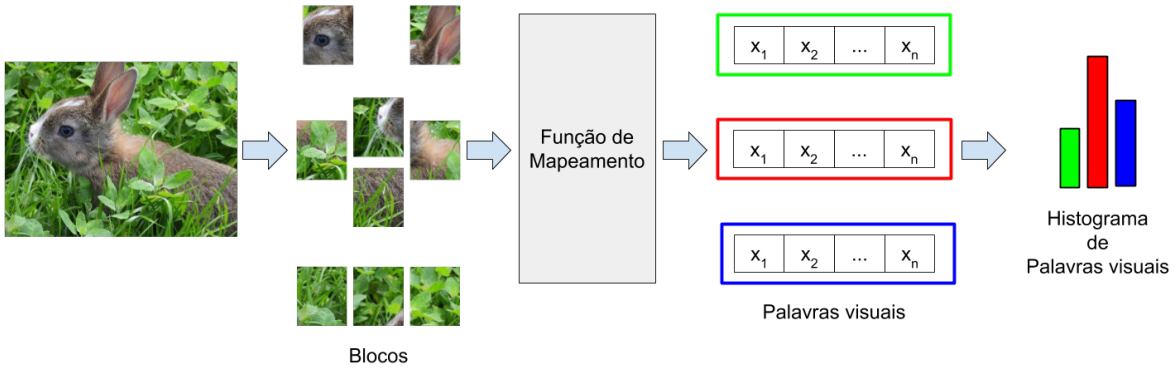
Fonte: Elaborada pelo autor.

Existem outras abordagens para realizar a associação do pontos de interesse às palavras visuais, essas abordagens levam em consideração a possibilidade de que um único ponto de interesse pode ser representado por mais de uma palavra visual. Philbin *et al.* (2008), Jiang, Ngo e Yang (2007) utilizam um tipo de abordagem classificada como *soft assignment* em que cada ponto de interesse pode ser associado a mais de uma palavra visual por meio de pesos. Já Jegou, Harzallah e Schmid (2007) utilizam em seu trabalho uma variação, que recebe o nome de *multiple assignment*, no qual os m vizinhos mais próximos são atribuídos como palavras visuais para cada ponto de interesse.

2.4 Discussão sobre a abordagem *Signature-based Bag of Visual Words*

O paradigma *Signature-Based Bag of Visual Words* (S-BoVW) foi proposto por Santos *et al.* (2015) e apresenta uma abordagem diferente da criação de dicionários de palavras visuais. Ao contrário dos métodos tradicionais, o S-BoVW não requer a etapa de agrupamento para definir as palavras visuais. Em vez disso, as imagens são divididas em blocos ou regiões de interesse, e cada bloco é mapeado diretamente em uma palavra visual por meio de uma função que representa suas características locais. Dessa forma, dado uma imagem I e um conjunto de blocos $B = b_1, b_2, ..., b_n$, é necessário apenas uma função F que mapeie cada bloco em uma palavra visual. A Figura 20 ilustra esse processo. Portanto, o principal fator que diferencia os métodos dentro do paradigma S-BoVW é a função de mapeamento utilizada para gerar as palavras visuais.

Figura 20 – Mapeamento dos blocos em palavras visuais.

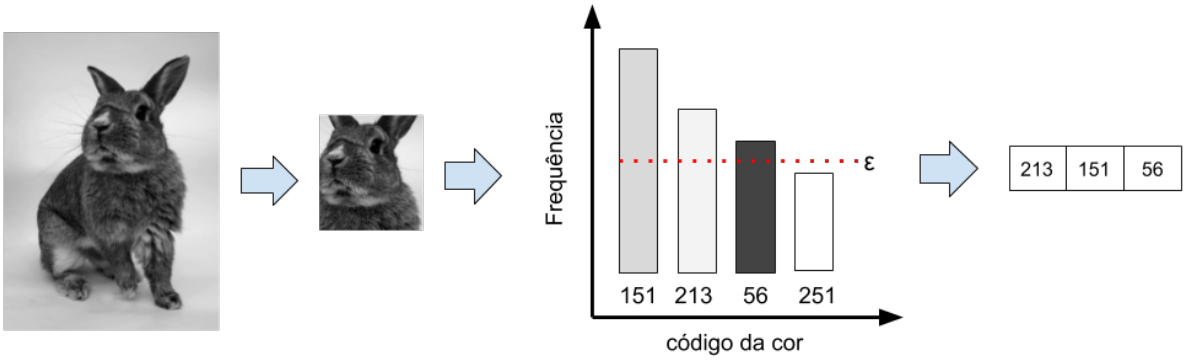


Fonte: Elaborada pelo autor.

Uma das primeiras funções que surgiram nesse paradigma foi a *Sorted Dominant Local Color* (SDLC) proposta por Santos (2016). Essa função gera uma assinatura textual para os blocos de uma imagem a partir de seu histograma de cores. O objetivo é que as cores mais frequentes possam representar o conteúdo do bloco. Para obter a assinatura, primeiro é gerado o histograma h_i para o bloco b_i. Em seguida, todas as cores que ocorrem menos do que ε vezes são removidas do histograma, onde ε é um limiar de frequência definido previamente. O resultado é um vetor com os códigos das cores resultantes ordenadas em ordem decrescente, que é utilizado para representar o bloco. A Figura 21 ilustra esse processo.

A *Sorted Dominant Local Texture* (SDLT) é uma função de mapeamento similar à SDLC, mas definida com o objetivo de representar a textura presente nos blocos. Para cada bloco b_i o código LBP é calculado para todos seus pixeis, e a partir disso é gerado um histograma de textura (contendo a frequência dos códigos LBP no bloco). Dessa forma a assinatura textual de cada bloco é gerada por meio dos códigos de textura mais frequentes, podando aqueles que

Figura 21 – Transformando um bloco em palavra visual por meio do SDLC.

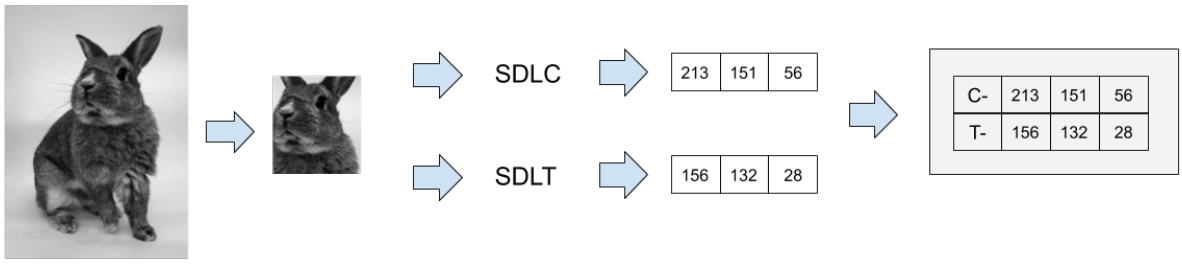


Fonte: Adaptada de Santos (2016, Página 28).

estiverem abaixo de um limiar ε de frequência e ordenando os códigos de textura resultantes em ordem decrescente.

Santos (2016) também apresenta o *Sorted Dominant Local Color and Texture* (SDLCT). No SDLCT, para gerar a assinatura textual dos blocos, são considerados os valores mais frequentes de cor e textura. Uma única assinatura é gerada a partir da combinação das funções SDLC e SDLT, de forma que as assinaturas geradas pelo SDLC recebem o prefixo "c-"para indicar que são assinaturas de cor, enquanto as assinaturas geradas pelo SDLT recebem o prefixo "t-". Assim as características de cor e textura se complementam, formando uma representação mais completa das imagens. A Figura 22 ilustra esse processo.

Figura 22 – Combinando o SDLC e SDLT para gerar uma palavra visual utilizando cor e textura.



Fonte: Adaptada de Santos (2016, Página 29).

Nessa seção, foi apresentado o paradigma S-BoVW, que propõe uma abordagem diferente do tradicional C-BoVW, no qual não há necessidade de construir um dicionário de palavras visuais. O S-BoVW permite definir palavras visuais a partir de funções que mapeiam o conteúdo de blocos em assinaturas textuais.

2.5 Considerações Finais

Neste capítulo, foram apresentadas as principais técnicas que embasam a pesquisa desenvolvida neste mestrado, concentrando-se nos conceitos básicos relacionados à representação de imagens em sistemas de recuperação de imagem por conteúdo. Através da exploração desses fundamentos, buscamos estabelecer uma base sólida para o desenvolvimento do projeto de mestrado e a busca por respostas às questões levantadas.

CAPÍTULO

3

TRABALHOS RELACIONADOS

3.1 Considerações Iniciais

Este capítulo tem como objetivo revisar a literatura sobre métodos desenvolvidos por outros autores que podem contribuir para o presente trabalho. A representação de características visuais de imagens é uma área de grande importância para sistemas CBIR. Serão apresentados métodos que abordam tanto as características globais quanto as locais. O foco dos estudos foram os métodos baseados em redes complexas e no modelo *Bag of Visual Words*. Ao revisar esses trabalhos, será possível identificar as contribuições e limitações de cada método, bem como como eles podem ser aplicados no contexto deste trabalho.

3.2 Abordagens baseadas em redes complexas

Dentre as abordagens utilizadas para caracterizar a textura, as redes complexas têm ganhado destaque na literatura. Essas redes permitem modelar a relação entre os elementos de uma imagem de forma precisa e eficiente, gerando representações que podem ser utilizadas em diversas aplicações. Nesta seção, serão apresentados alguns dos trabalhos que utilizam redes complexas para caracterizar a textura em imagens.

Tipicamente essas abordagens modelam a imagem I como um grafo G, onde cada pixel é um vértice e existem arestas entre pixeis com distância menor ou igual a um raio r previamente definido. O peso da aresta normalmente é definido por uma função que leva em consideração a distância entre os pixeis e suas diferenças de intensidade.

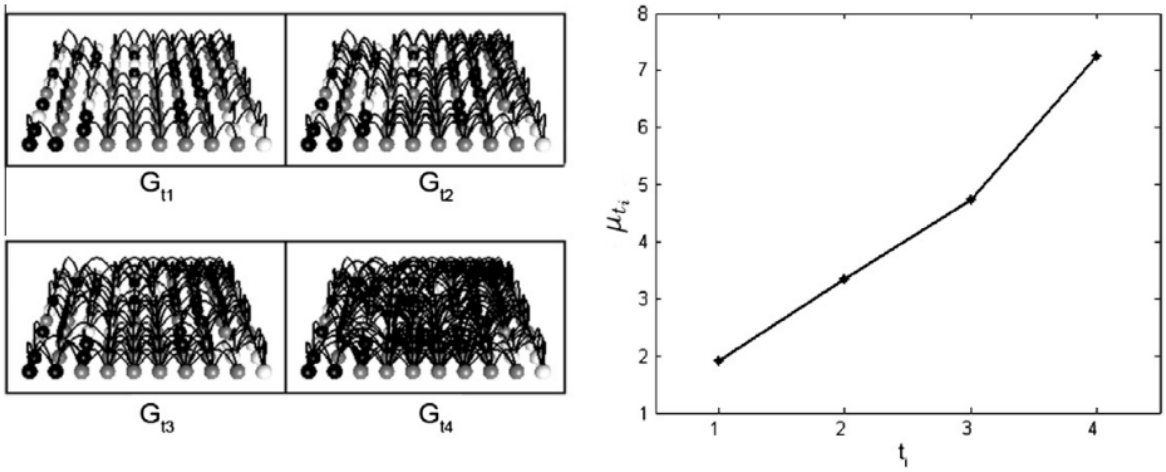
Observe que, dessa forma, cada vértice da rede apresenta um número semelhante de conexões, gerando um grafo regular que não apresenta propriedades relevantes para a análise das características visuais da imagem. Assim, é necessário transformar esse grafo em uma rede complexa que possua características relevantes para as análises. Isso normalmente é feito

aplicando um conjunto $T = t_1, t_2, ..., t_n$ de limiares, de forma que são gerados subgrafos G_t onde existem apenas as arestas onde o peso é menor ou igual à t.

O trabalho apresentado em (BACKES; CASANOVA; BRUNO, 2013), propôs um método de análise de textura que calcula medidas estatísticas como a média, contraste, energia e entropia a partir do histograma de grau dos subgrafos G_t gerados ao aplicar os limiares na etapa de transformação. Gerando um único vetor de características que contém as medidas concatenadas de cada uma das redes e irá representar a imagem.

A análise de cada uma das redes geradas permite observar suas características topológicas, como o grau médio que varia de acordo com o limiar aplicado como ilustrado pela Figura 23.

Figura 23 – A figura mostra a evolução de uma rede com base no limiar aplicado, como podemos ver o grau médio da rede e suas características topológicas variam conforme o liminar t_m aplicado



Fonte: Backes, Casanova e Bruno (2013, Página 171).

Visando ampliar as estratégias de representação de imagens com redes complexas (CANTERO *et al.*, 2020) propôs um método baseado na importância dos vértices, que pode ser descrito em três etapas: modelagem da imagem como uma rede complexa; extração da importância de cada vértice usando *pagerank* (PAGE *et al.*, 1999) e cálculo do vetor de características que correlaciona a importância dos vértices e seus graus para descrever a imagem.

Embora o *pagerank* tenha sido projetado para classificar páginas da Web, também pode ser aplicado a grafos em geral. Neste caso, o *pagerank* atribui uma pontuação para descrever a importância de cada vértice com base em suas conexões com os vizinhos.

Para montar a representação final da imagem é calculado um histograma que correlaciona a importância $\rho(v)$ de um vértice $v \in V$ com seu grau $k(v)$ como ilustrado pela Equação 3.1, nessa equação j indica o *bin* do histograma e varia até o grau máximo da rede.

$$h_{r,t}(j) = \sum_{v \in V} \begin{cases} \rho_{r,t}(v), & \text{Se k(v) = j} \\ 0, & \text{caso contrário} \end{cases} \tag{3.1}$$

A maioria dos métodos discutidos até agora consideram apenas um canal de cores da imagem, normalmente considerando apenas seus valores de intensidade. O trabalho apresentado em (SCABINI *et al.*, 2019) propõe uma nova técnica para modelar e caracterizar cor e textura em imagens utilizando redes complexas. Essa abordagem considera cada canal de cor como uma camada da rede, onde os pixeis da imagem são representados pelos vértices. São aplicadas transformações na rede resultante com base em limiares previamente definidos para capturar informações de cor e textura em diferentes momentos da evolução da rede .

Para modelar a imagem como uma rede, os autores consideram uma imagem I com N pixeis e Z canais de cores, onde o número total de vértices da rede seria N * Z. As arestas são adicionadas entre os pixeis que possuem distância menor ou igual a um raio r pré-definido. Em seguida, as redes resultantes são caracterizadas utilizando medidas estatísticas, como média, desvio padrão, entropia e energia, baseadas na distribuição de grau e coeficiente de agrupamento em cada uma das redes resultantes de acordo com os limiares aplicados. Assim como ilustrado pela Figura 24

Os resultados do trabalho mostram que as redes que combinam as informações dos canais de cores fornecem características ricas e discriminantes para a maioria das bases de dados analisadas e são mais eficientes que métodos baseados somente nos níveis de intensidade das imagens. Com isso, essa técnica proposta pode ser útil em diversas aplicações, como reconhecimento de padrões e análise de imagens médicas, entre outras.

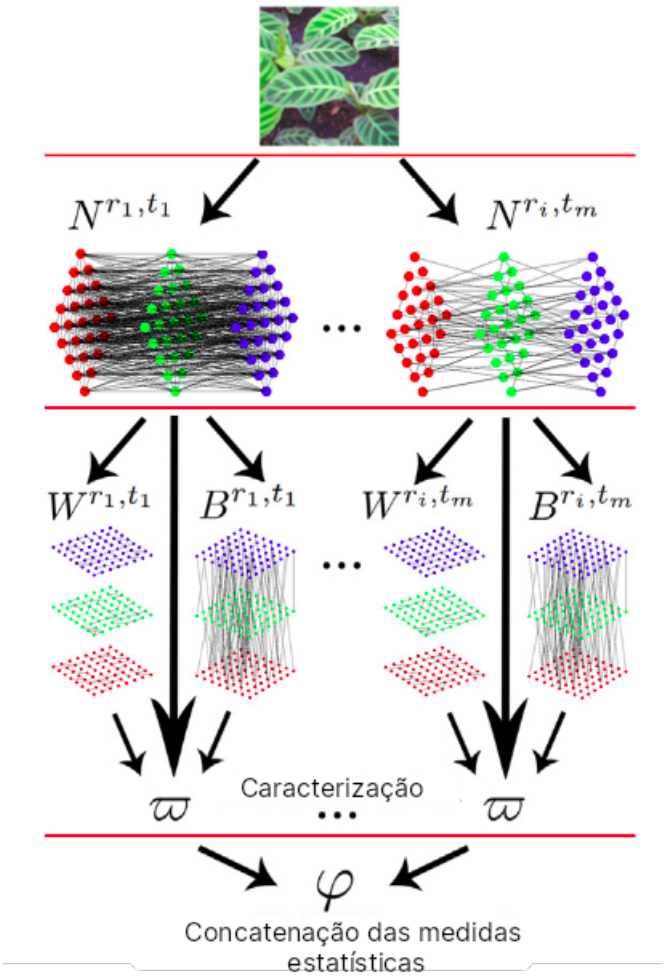
Todos os métodos citados até aqui obtiveram bons resultados quando comparado com descritores de textura tradicionais. No entanto, possuem alguns problemas em comum. Eles constroem inicialmente um único grafo para representar a imagem como um todo, o que é um processo caro em termos de custo computacional e pode não ser eficiente em imagens com estruturas complexas e heterogêneas, podendo levar a perda de informações importantes e afetar negativamente a performance do método.

3.3 Abordagens baseadas no paradigma C-BoVW

As técnicas C-BoVW são usadas para extrair características visuais de uma imagem e representá-las em forma de histograma de palavras visuais, onde cada palavra visual é um *cluster* de características visuais semelhantes. Essa representação baseada em histogramas por padrão não considera informações de localização espacial das palavras visuais pois tratam cada palavra visual como uma entidade independente, sem considerar sua posição espacial na imagem. A seguir serão apresentados alguns trabalhos que tentam aprimorar a abordagem tradicional C-BoVW com foco em sua representação espacial.

O trabalho apresentado em Pedrosa (2015) propõe o *Bag-of-2-Grams* que utiliza frases visuais para modelar a coocorrência das palavras visuais e capturar relação espacial entre as palavras visuais. Dessa forma, ao invés de se utilizar um dicionário de palavras, o autor utiliza

Figura 24 – Passo a passo para a aplicação do método, primeiro a imagem é modelada como um grafo, os limiares são aplicados parar gerar sub-redes e a caracterização é feita com aplicando medias estatísticas nas distribuições de grau e agrupamento das redes resultantes



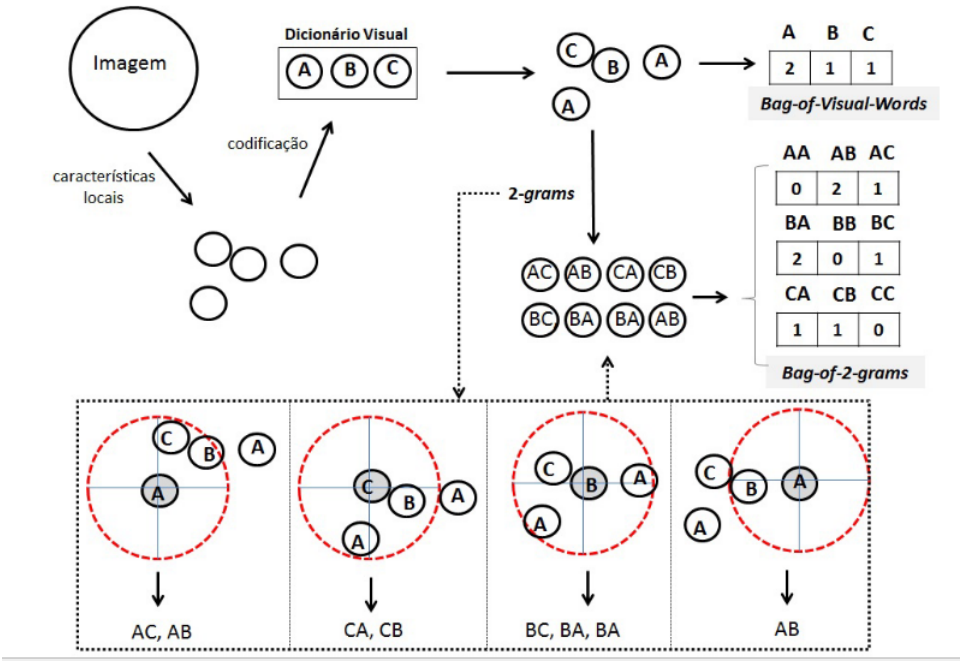
Fonte: Scabini *et al.* (2019, Página 40).

um dicionário de frases. Para atingir esse objetivo o autor propõe o uso de frases de tamanho fixo, também conhecidas como *n-grams*, de forma que n é o número de palavras da frase, no entanto isso pode levar a um dicionário de frases visuais muito grande.

Para resolver esse problema o autor trabalha com *2-grams*. Os *2-gram* são definidos como todos os pares de palavras visuais dentro de uma vizinhança com raio r, e por fim a representação final da imagem seria dada pela frequência dos *2-grams* encontrados em cada imagem. A Figura 25 ilustra esse processo.

O *Global Spatial Arrangement* (GSA), desenvolvido por Pedrosa (2015), trabalha com a descrição de como as palavras visuais estão globalmente distribuídas nas imagens. Dessa forma é possível saber em qual região uma palavra visual está mais presente na imagem, acima, abaixo, esquerda ou direita. A ideia central do trabalho consiste em, para cada ponto de interesse p_i montar um quadrante Q com quatro regiões $Q = \{Q1, Q2, Q3, Q4\}$ de forma que p_i represente o seu centro, $Q1$ a região superior esquerda, $Q2$ a região superior direita, $Q3$ a região inferior

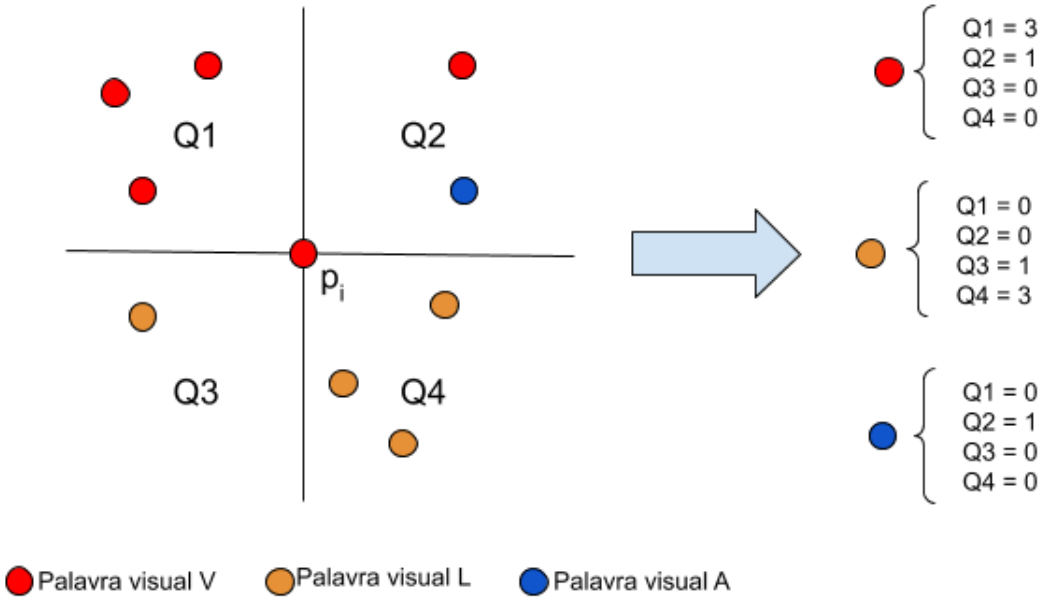
Figura 25 – Passo a passo para a geração do dicionário de frases utilizando o Bag-of-2-Grams.



Fonte: Pedrosa (2015, Página 53).

esquerda e $Q4$ a região inferior direita. Uma vez definido o quadrante Q é possível contar a ocorrência de cada palavra visual em cada uma das regiões de Q. Essa contagem é utilizada para obter a estimativa que indica em qual região da imagem uma palavra visual ocorre mais vezes. A Figura 26 ilustra o processo de contagem para um ponto p_i.

Figura 26 – Contagem de ocorrências das palavras visuais para um ponto de interesse p_i

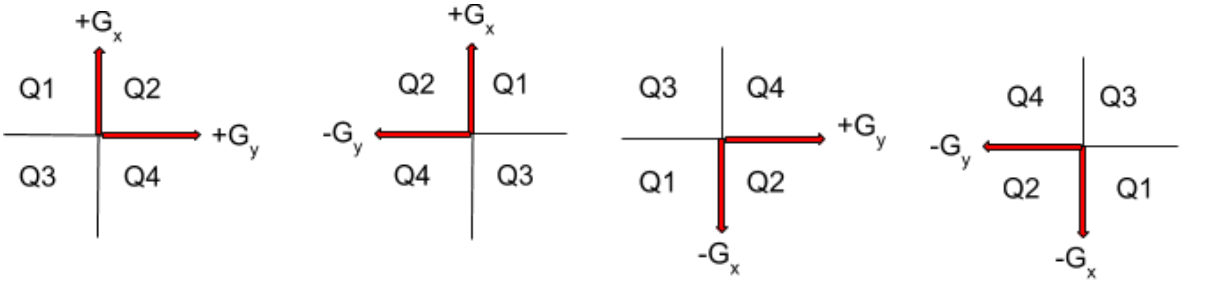


Fonte: Adaptada de Pedrosa (2015, Página 54).

Ao realizar a contagem descrita acima para todos os pontos de interesse em uma imagem

I, cada palavra visual seria representada por 4 valores, que indicam como ela está globalmente distribuída na imagem. No entanto, ao se aplicar o mesmo processo de contagem em uma imagem rotacionada I', teríamos uma contagem diferente para as palavras visuais. O autor propôs uma variante dessa abordagem para resolver o problema de invariância à rotação, na qual é utilizada a informação do gradiente no ponto de interesse p_i, para obter a disposição do quadrante Q. A direção do gradiente sempre aponta na direção de maior variação de intensidade, essa informação é utilizada para obter a posição de cada uma das regiões do quadrante Q sem se preocupar com rotações na imagem. A Figura 27 ilustra esse processo.

Figura 27 – Disposição das regiões do quadrante de acordo com a direção do gradiente. Considere a direção do gradiente de um ponto de interesse p_i como um vetor $G(p_i) = \{G_x(p_i), G_y(p_i)\}$



(a) Se $G_x(p_i) > 0$ $G_y(p_i) > 0$ (b) Se $G_x(p_i) > 0$ $G_y(p_i) < 0$ (c) Se $G_x(p_i) < 0$ $G_y(p_i) > 0$ (d) Se $G_x(p_i) < 0$ $G_y(p_i) < 0$

Fonte: Adaptada de Pedrosa (2015, Página 57).

Como discutido acima, essa estratégia gera 4 valores para representar cada palavra visual, dessa forma a representação final da imagem teria dimensionalidade $4K$, no qual K é o número de palavras visuais. Para contornar esse problema o autor propôs uma estratégia que resume esses valores em apenas dois, fazendo com que o vetor que representa a imagem passe de uma dimensionalidade de $4K$ para $2K$. Para atingir esse objetivo, cada região é sumarizada de acordo com sua posição no quadrante, top = Q1+Q2, left = Q1+Q3, right = Q2+Q4 e down = Q3+Q4, e por fim a representação final de cada palavra visual é dada por dois valores S_t e S_l como visto na Equação 3.2 e Equação 3.3.

$$S_{td} = \frac{top}{top + down} \tag{3.2}$$

$$S_{lr} = \frac{left}{left + right} \tag{3.3}$$

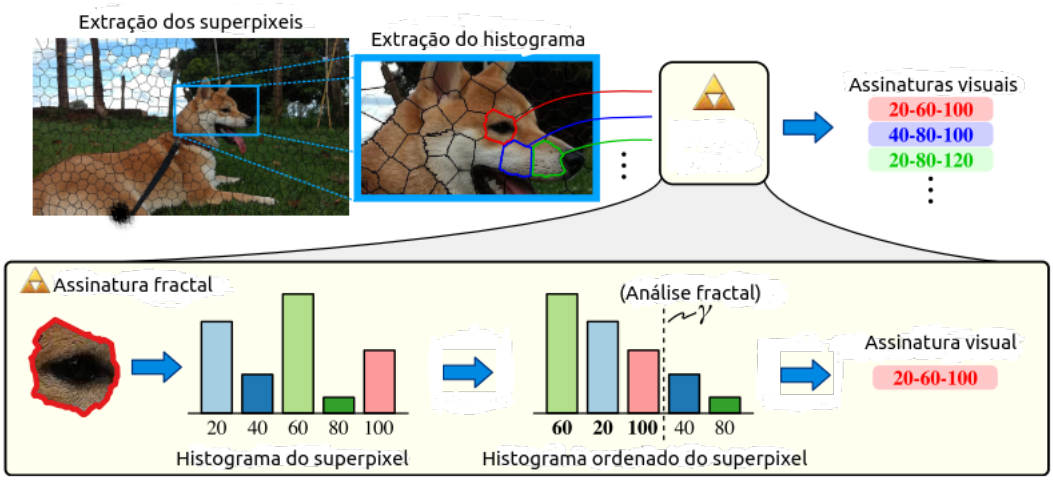
3.4 Abordagens baseadas no paradigma S-BoVW

O paradigma Signature-Based Bag of Visual Words (S-BoVW) apresentado na Seção 2.4 utiliza uma abordagem diferente da criação de dicionários de palavras visuais. Em vez disso, as

imagens são divididas em blocos ou regiões de interesse, e cada bloco é mapeado diretamente em uma palavra visual por meio de uma função que representa suas características locais.

Em Chino *et al.* (2018), os autores propuseram o Bag-Of-SuperPixel Signatures (BOSS), um método pertencente ao paradigma S-BoVW baseado em *SuperPixels* e cores dominantes, no qual o único parâmetro requerido é o número m de assinaturas visuais que serão extraídas da imagem. O método se baseia em três passos: a detecção de regiões, extração de características e por último a geração de assinaturas. Seja I uma imagem, o primeiro passo é utilizar um algoritmo de *superpixels* para gerar uma lista B = $\{b_1, b_2, ..., b_m\}$ de m blocos que serão convertidos em assinaturas visuais. No segundo passo para cada bloco b_i é gerado um histograma H que contabiliza a frequência das cores no bloco, esse histograma posteriormente é ordenado de forma decrescente. No terceiro passo, cada histograma é passado para o módulo fractal que retorna somente os γ valores mais frequentes para gerar a assinatura visual final de cada bloco (superpixel). Assim como ilustrado pela Figura 28.

Figura 28 – Transformando um bloco em palavra visual por meio do BOSS



Fonte: Adaptada de Chino *et al.* (2018, Página 66).

Um fator importante a se destacar é como o método determina o valor γ, esse processo é feito mediante o cálculo da dimensão intrínseca da região da imagem, ou dimensão fractal, que define o número mínimo de atributos necessários para representar um ponto em um conjunto de dados (JR; TRAINA; FALOUTSOS, 2010). Considera-se o conjunto de todos os histogramas ordenados pela frequência, isto é em ordem descendente como Z em que cada elemento tem q atributos, ao calcular a dimensão intrínseca deste conjunto, temos o valor mínimo de atributos necessários para representar um elemento pertencente a Z. Portanto um bom valor para o parâmetro γ seria pelo menos a dimensão intrínseca de Z.

Os autores ainda propõem algumas variações, a primeira delas o BoSS-T representa as imagens por meio de assinaturas visuais de textura. Cada pixel da imagem é descrito por um valor de textura, o método utilizado pelos autores para essa tarefa foi o LBP. Dessa forma é possível criar um histograma de texturas que irá representar cada bloco da imagem pelos seus

valores de textura. Por fim o histograma é ordenado de forma decrescente e podado utilizando o parâmetro γ para gerar a assinatura visual.

A segunda variação se chama BOSS-CT e combina as duas abordagens citadas anteriormente, após obter as assinaturas de cor e textura, cada uma delas recebe um prefixo. o prefixo C é utilizando para as assinaturas de cor e T para as assinaturas de textura. Após isso a representação final das imagens pode ser construída considerando a frequência das assinaturas visuais encontradas.

3.5 Abordagens baseadas na combinação do BoVW e redes complexas

As abordagens que combinam o modelo *Bag of Visual Words* com redes complexas foram pouco exploradas na literatura. No entanto, alguns estudos já mostraram que essa é uma estratégia promissora e pode ser explorada para gerar representações poderosas das imagens. Essas abordagens pode permitir a extração de características visuais globais e locais da imagem, levando a uma representação mais completa e discriminativa das características da imagem.

Em (SCABINI; GONÇALVES; CASTRO, 2015), os autores propuseram um método para análise de texturas que combina o *Bag of Visual Words* com redes complexas. O método segue o mesma modelagem citada anteriormente na Seção 3.2 para transformar uma imagem I em grafo G e para a geração de redes intermediárias G_t com base no limiar $t \in T$ aplicado.

Resumidamente, em cada rede G_t com arestas entre vértices cuja distância é menor ou igual a um raio r, cada vértice v_i é representado por um vetor de características $[k^{t1,r}, s^{t1,r}, k^{t2,r}, s^{t2,r}, ..., k^{tn,r}, s^{tn,r}]$. O valor $k^{t,r}$ indica o grau de v_i na rede resultante da aplicação do limiar t e do raio r, enquanto $s^{t,r}$ representa a força do vértice nessas mesmas condições.

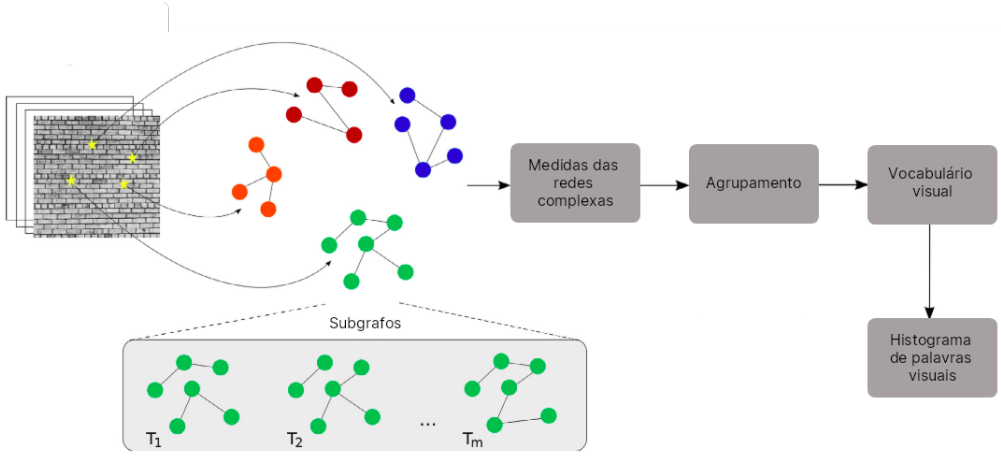
Após a criação dos vetores de características que representam cada vértice (pixel), o BoVW pode ser utilizado para gerar o dicionário de palavras visuais a partir das representações obtidas. Em seguida, cada palavra visual é atribuída a um vértice com base na distância euclidiana entre a representação do vértice e as palavras visuais do dicionário. Por fim, o histograma resultante que representa a imagem, é construído.

O método proposto por Lima *et al.* (2019) apresenta uma nova abordagem denominada BoVW-CN, aqui temos a utilização de redes complexas para extrair as características de texturas de regiões de interesse na imagem. O Primeiro passo do método é construir o dicionário de palavras visuais. Para isso é utilizado um detector de ponto de interesses que terá como resposta um conjunto $P = \{p_1, p_2, ..., p_n\}$. A partir disso um grafo G_i é montando para cada p_i, utilizando os pixeis de sua vizinhança.

Para as análises comportamentais dos grafos, os autores fazem análises de seus subgrafos aplicando um conjunto $T = \{t_1, t_2, ..., t_n t\}$ de limiares, de forma que ao aplicar o limiar t somente

restarão no grafo as arestas com peso menor ou igual à t. Dessa forma para cada grafo G_{i_t} são aplicadas um conjunto de características para analisar sua topologia e gerar um vetor que irá representar suas características de textura. Para construir o vocabulário de palavras visuais, os autores utilizaram o algoritmo de agrupamento *k-means*, selecionando os centroides como palavras visuais. A função de distância adotada foi a distância Euclidiana. Por fim, para finalizar a representação das imagens é criado um histograma com as palavras visuais. Assim como ilustrado pela Figura 29.

Figura 29 – Aplicação do método BoVW-CN, primeiro são detectadas regiões de interesse nas imagens e a partir disso serão construídas redes complexas utilizando a vizinhança de cada uma dessas regiões. Medidas baseadas na teoria dos grafos são aplicadas para representar cada região de interesse que posteriormente são agrupadas para criar o vocabulários visual e a representação das imagens.



Fonte: Adaptada de Lima *et al.* (2019, Página 217).

Comparando o método com técnicas tradicionais que utilizam redes complexas para extrair características de textura, o custo computacional do BoVW-CN é baixo, pois as redes complexas são criadas a partir da vizinhança dos pontos de interesse ao invés de se considerar toda a imagem. No entanto, para se ter bons resultados na representação gerada para as imagens é preciso ter cuidado na escolha do método de detecção dos pontos de interesse, pois segundos os autores ele tem forte influência no resultado final.

3.6 Considerações Finais

Após a revisão dos trabalhos relacionados, ficou evidente que a análise de características visuais de imagens é um campo de pesquisa em constante evolução, com uma grande variedade de abordagens e técnicas propostas nos últimos anos. Embora a maioria dessas abordagens tenha sido bem-sucedida em extrair informações relevantes das imagens, ainda há limitações a serem superadas. Em resumo, a revisão dos trabalhos relacionados destaca a importância da escolha da abordagem mais adequada para a análise de características visuais de imagens, com base nas limitações e benefícios de cada método.

CAPÍTULO

4

BOCS: *BAG OF COMPLEX SIGNATURES*

4.1 Visão Geral

A representação de imagens é um desafio para os sistemas CBIR, pois exige a capacidade de identificar características visuais relevantes e transformá-las em uma representação que possa ser usada para recuperar imagens semelhantes. O objetivo principal deste trabalho é desenvolver uma técnica de representação de imagens adequada para sistemas de recuperação de imagens por conteúdo, que possa proporcionar resultados de busca mais precisos e de acordo com as expectativas do usuário.

A técnica conhecida como *Bag of Visual Words* é uma ferramenta amplamente utilizada para essa tarefa, pois permite a descrição das imagens por meio de assinaturas visuais, que podem preservar o conteúdo semântico das imagens se forem selecionadas adequadamente. Como discutido nas Seção 2.2 e Seção 3.2, as abordagens baseadas em redes complexas têm se mostrado bem-sucedidas na representação de imagens para diversas aplicações. Portanto esse tipo de técnica se torna uma candidata em potencial para descrever regiões de interesse na imagem como assinaturas visuais preservando seu conteúdo semântico.

Entretanto, a combinação dessas duas técnicas tem sido pouco explorada na literatura. Logo, este trabalho propõe o método *Bag of Complex Signatures* (BoCS) que une ambas as técnicas em uma abordagem baseada no paradigma S-BoVW (discutido na Seção 2.4), visando melhorar a representação semântica das imagens, o que pode contribuir para uma busca mais efetiva em sistemas de recuperação de imagens por conteúdo.

4.2 Metodologia

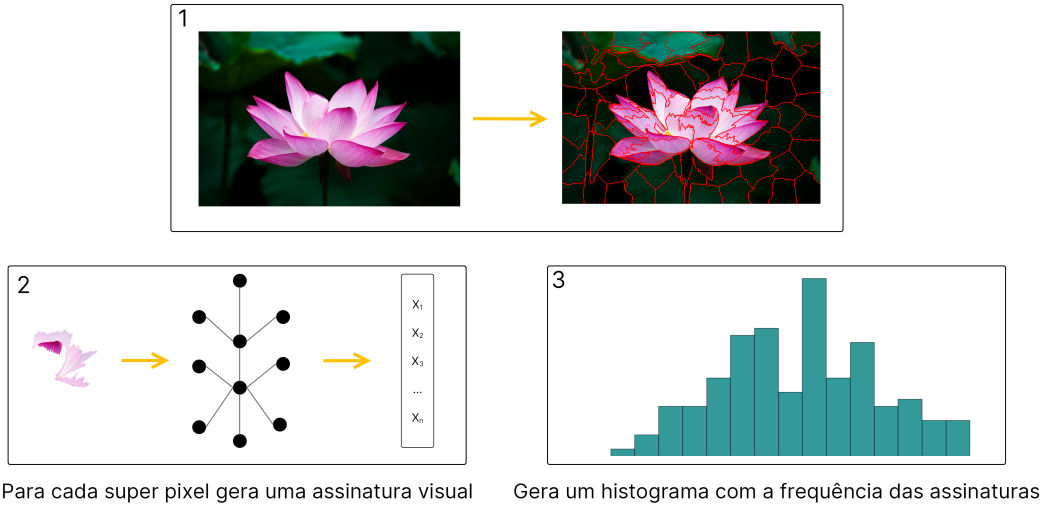
A técnica proposta consiste em três etapas. Na primeira etapa, a imagem é dividida em pequenas regiões chamadas super pixeis, que são conjuntos de pixeis com similaridade em

termos de cor e textura. Isso permite a criação de uma representação mais compacta da imagem, com foco nas regiões de maior interesse sem perda de informações importantes.

Na segunda etapa, cada super pixel é transformado em uma rede complexa, onde os pixeis são considerados como nós da rede e as conexões entre eles são definidas com base na proximidade espacial e na diferença de seus níveis de intensidade. Essa rede é utilizada para extrair características de textura da imagem, permitindo uma descrição dos padrões visuais presentes na imagem. Essa descrição extraída será utilizada como assinatura visual.

Por fim, na terceira etapa, a representação da imagem é construída com base na frequência de ocorrência das assinaturas visuais dos super pixeis, resultando em um histograma que conta quantas vezes cada assinatura visual aparece na imagem assim como ilustrado pela Figura 30. Nas seções a seguir serão detalhadas cada uma destas etapas.

Figura 30 – Visão geral da metodologia proposta dividia em três etapas



Fonte: Elaborada pelo autor.

4.2.1 Divisão em blocos

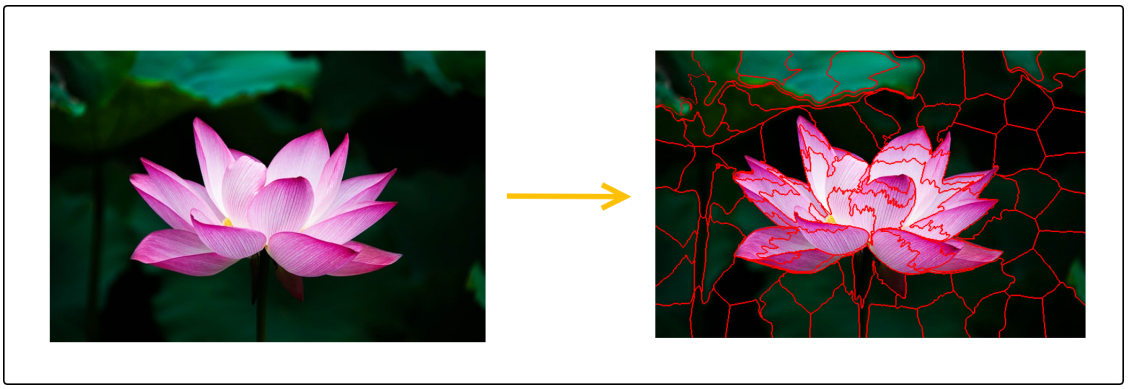
O primeiro passo do BoCS é dividir a imagem em regiões de interesse. Neste trabalho utilizamos os algoritmos de super pixeis, essa é uma técnica de processamento de imagens que agrupa pixeis de uma imagem em regiões semelhantes capturando redundâncias. Dessa forma ao invés de trabalhar com cada pixel individualmente, a imagem é dividida em regiões maiores, normalmente utilizando um algoritmo que leva em conta a similaridade entre os pixeis observando sua cor, textura e intensidade.

A Figura 31 ilustra o processo de divisão da imagem em um conjunto $B = b_1, b_2, ..., b_n$ de blocos. Essa abordagem simplifica a construção de assinaturas visuais, já que cada super pixel pode ser transformado em uma assinatura visual e se a representação gerada for geral o suficiente, super pixeis similares podem ser mapeados para uma única assinatura.

Existem diversos algoritmos para a geração de super pixeis, cada um com suas vantagens, desvantagens e que podem ser utilizados em aplicações específicas. Por conta disso os autores em (STUTZ; HERMANS; LEIBE, 2018) realizaram uma ampla comparação entre os métodos disponíveis na literatura e os classificaram. Definindo aqueles que são os algoritmos mais recomendados para a utilização em aplicações gerais, eles apontam o *Simple Linear Iterative Clustering* (SLIC) (ACHANTA *et al.*, 2010) como um dos métodos mais generalistas.

O SLIC começa com uma inicialização aleatória dos centros dos super pixeis, que posteriormente são ajustados iterativamente para minimizar uma função de custo que combina a distância de cor e a distância espacial entre os pixeis e o centro de super pixel mais próximo, além disso sua execução é rápida e eficiente em termos de tempo e ainda produz regiões com boa aderência às bordas e formato regular, mostrando-se uma boa opção para ser utilizada durante os experimentos deste trabalho.

Figura 31 – Identificando regiões de interesse na imagem utilizando super pixeis



Fonte: Elaborada pelo autor.

4.2.2 Modelagem dos blocos como grafos

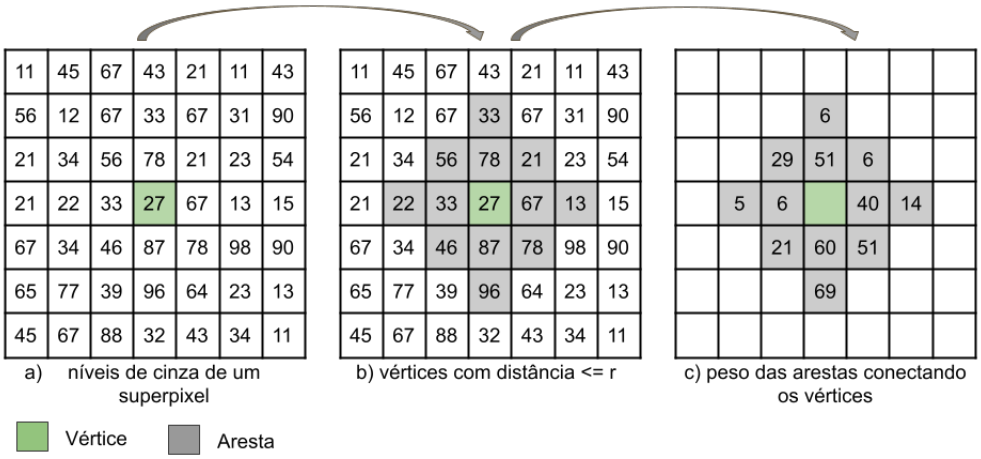
Modelar cada super pixel como um grafo é um dos pontos principais para analisar a textura com o BoCS. Super pixeis são regiões da imagem que foram agrupadas em uma única entidade, o que nos permite gerar grafos menores quando comparado com abordagens que transformam toda uma imagem em um único grafo. Isso torna a análise mais eficiente em termos computacionais, já que podemos analisar cada região da imagem separadamente. Ao modelar um super pixel como um grafo, podemos usar as arestas para capturar informações sobre a diferença de intensidade entre os vértices(pixeis). O peso das arestas pode ser definido de várias maneiras, mas uma abordagem comum é usar uma função que leva em conta a diferença de intensidades e a distância entre os vértices.

Portanto para esta etapa de modelagem, construímos para cada bloco $b_i \in B$, um grafo g_i de forma que cada pixel seja um vértice $v_i \in V$. Dois vértices v_i e v_j apenas estão conectados por uma aresta $e_{v_i,v_j} \in E$ se a distância entre eles for menor ou igual a um determinado raio r. O peso w é definido observando a diferença de intensidades e a distância entre os pixeis, normalizados

pelo raio r. Isso faz com que o peso máximo possível para uma aresta seja igual ao maior valor de intensidade na imagem. Assim, o modelo captura a relação espacial entre os pixeis da imagem. Além disso, a normalização do peso pelo raio r garante que a distância não tenha um efeito desproporcional em relação às diferenças de intensidades entre os pixeis.

A Figura 32 traz um exemplo onde criamos as conexões de um pixel (destacado em verde) com seus vizinhos (destacados em cinza) e Equação 4.1 traz mais detalhes sobre a função de peso utilizada.

Figura 32 – a) Cada pixel do bloco é um vértice no grafo, o vértice em verde está em destaque para analisarmos seus vizinhos; b) Dois vértices estão conectados se a distância Manhattan entre eles é menor que um raio r (conexões destacadas em cinza); c) o peso das arestas é calculado observando a distância e a diferença de intensidade entre os vértices



Fonte: Elaborada pelo autor.

Usando essa função de peso, podemos usar algoritmos de processamento de grafos para extrair informações relevantes para a análise da textura, um exemplo disso é a distribuição da força dos vértices, pelo seu formato podemos identificar a presença de regiões da imagem com texturas fortes e intensidades distintas(grafos onde o peso das arestas possuem em sua maioria valores altos, o que indica alta variação de intensidade), ou a presença de regiões mais homogêneas na imagem(grafos com arestas com pesos baixos, o que indica regiões com baixa diferença de intensidade).

$$w(e_{v_i,v_j}) = \frac{|x_{v_i} - x_{v_j}| + |y_{v_i} - y_{v_j}| * |I(v_i) - I(v_j)|}{r} \tag{4.1}$$

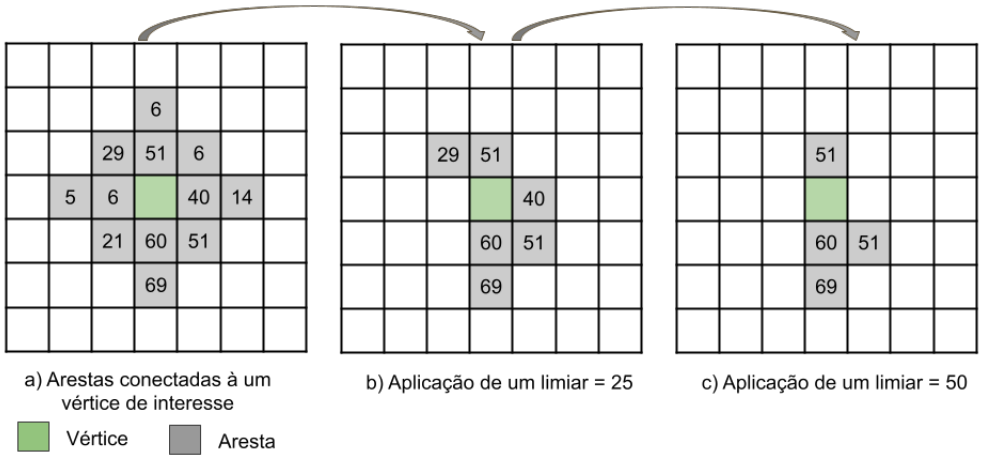
4.2.3 Transformando os grafos em redes complexas

O grafo que construímos na Seção anterior é regular pois as arestas são definidas com base em uma distância r e portanto todos os seus vértices tem o mesmo número de conexões (exceto os vértices que estão nas bordas). Logo é importante aplicar transformações em sua estrutura para que possamos computar propriedades das redes obtidas.

Na literatura, uma das transformação mais comuns a ser aplicada para esse tipo de problema é definida como *thresholding*, onde podamos do grafo arestas que estão fora de um determinado conjunto $T = \{t_1, t_2, ..., t_n\}$ de limiares, dessa forma é possível gerar redes complexas com características distintas de acordo com o limiar aplicado, uma vez que estaremos retirando arestas dos vértices como ilustrado pela Figura 33 e por consequência alterando toda a estrutura da rede.

Esse processo pode ser interpretado como a aquisição de várias amostras de redes e nos dá a capacidade de gerar uma representação com um rico conjunto de características que irão descrever o seu comportamento. Neste trabalho estamos removendo do grafo todas as arestas que são menores do que o limiar aplicado. Portanto ao aplicar um limiar na rede, destacamos as conexões onde ocorrem as maiores diferenças de intensidade, pois removemos arestas com menor peso (ou seja as conexões que representam uma baixa diferença de intensidade) e destacamos aquelas onde há uma maior diferença de intensidade. Como texturas são caracterizadas por variações de intensidade, esse processo nos ajuda a destacar tipos de texturas na imagem de acordo com o limiar aplicado.

Figura 33 – a) Mostra um vértice destacado em verde e suas arestas destacadas em cinza, com seus respectivos pesos; b) Mostra a aplicação de um limiar = 50 eliminando arestas que estão abaixo dele; c) Aplicação de limitar = 25 eliminando arestas que estão abaixo dele



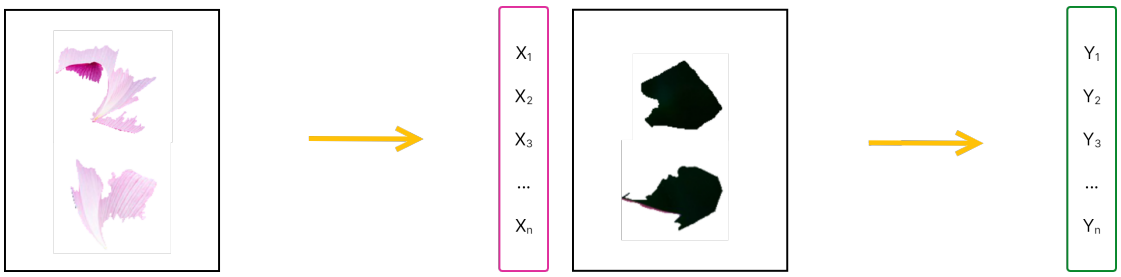
Fonte: Elaborada pelo autor.

4.2.4 Gerando as assinaturas visuais

Após aplicar o *thresholding* é possível estudar a estrutura das redes geradas, podendo utilizar ferramentas já bem estabelecidas na teoria dos grafos para analisar a textura da imagem. Uma das ferramentas mais comuns é a análise da distribuição de grau dos vértices. No entanto essa abordagem é limitada, pois ela não traz informações diretas sobre a intensidade dos pixeis. Portanto podemos utilizar a distribuição de força dos vértices, pois ela nos fornece informações mais diretas sobre a variação de intensidade na região analisada.

Um ponto importante em métodos baseados no paradigma S-BoVW é que suas funções de mapeamento precisam ser genéricas o suficiente para representar o conteúdo de regiões similares como uma única assinatura visual e ainda assim conseguir que regiões diferentes tenham assinaturas distintas (SANTOS *et al.*, 2015), assim como ilustrado pela Figura 34.

Figura 34 – Pode-se ver pela figura que a ideia é identificar regiões similares e mapeá-las em assinaturas únicas, em (a) as pétalas que possuem características visuais semelhantes podem ser mapeadas para uma mesma representação, enquanto as folhas de fundo na parte (b) podem ser descritas com uma nova representação.



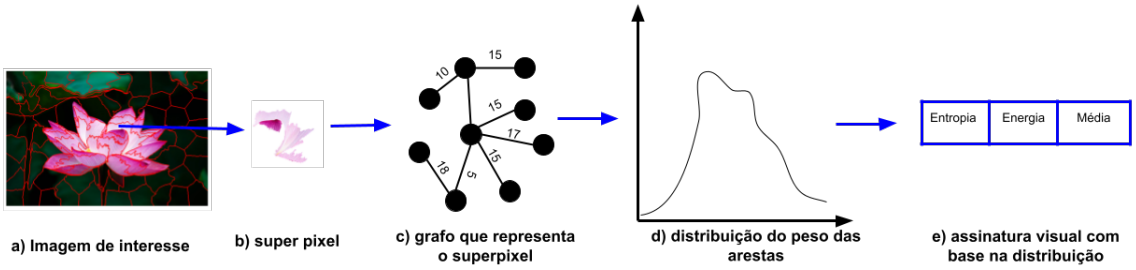
(a) Pétalas sendo mapeadas em uma única assinatura visual

(b) Folhas de fundo sendo mapeados em uma única assinatura visual

Fonte: Elaborada pelo autor.

Se utilizarmos uma distribuição como assinatura visual uma pequena alteração na intensidade de pelo menos um de seus pixeis poderia gerar uma nova assinatura visual, dessa forma teríamos uma função de mapeamento muito específica e sem capacidade de generalização. Portanto, para gerar a assinatura que irá representar cada bloco da imagem precisamos resumir a informação presente em cada rede e capturar comportamentos gerais que não mudem conforme mudanças pequenas ocorrerem na topologia da rede. Para esse passo podemos utilizar estatísticas como média, energia e entropia para resumir o conteúdo da distribuição analisada. A Figura 35 ilustra esse processo.

Figura 35 – a) Imagem de interesse; b) super pixel b_i que está sendo analisado; c) grafo g_i construído para representar o super pixel; d) distribuição dos pesos das arestas; e) medidas retiradas da distribuição permitindo representar o bloco como uma assinatura visual



Fonte: Elaborada pelo autor.

No contexto da distribuição de força, a média representa a diferença de intensidade média das conexões entre os pixeis. Uma média alta pode indicar a presença de texturas mais fortes e contrastantes, enquanto uma média baixa pode indicar regiões mais homogêneas. A entropia

mede o grau de aleatoriedade da distribuição de força das arestas. Quanto maior a entropia, maior é a diversidade das conexões presentes na rede. Uma alta entropia pode indicar a presença de texturas mais complexas e variadas, enquanto uma entropia baixa pode indicar regiões mais uniformes e simples. A energia representa a soma dos quadrados da distribuição de força das arestas, dando uma ideia da magnitude total das conexões presentes na rede. Uma alta energia pode indicar a presença de texturas mais densas e com conexões mais fortes, enquanto uma energia baixa pode indicar regiões mais dispersas e com conexões mais fracas.

- Média:
$$\mu = \frac{1}{N}\sum_{i=1}^{N} w_i$$
- Entropia:
$$\phi = -\sum_{i=1}^{N} p_i \log_2 p_i$$
onde p_i é a probabilidade de ocorrência do peso w_i.
- Energia:
$$\varepsilon = \sum_{i=1}^{N} w_i^2$$

Além da distribuição de força, podemos completar nossa análise estudando o formato das conexões na rede. O coeficiente de agrupamento é uma medida que fornece informações sobre a estrutura local de um grafo. Ele indica o quanto os vizinhos próximos a um determinado vértice estão conectados entre si. Essa medida é importante para entender como os pixeis de uma imagem estão conectados entre si em uma região específica. Ao utilizar a distribuição do coeficiente de agrupamento em uma imagem modelada como grafo, é possível identificar regiões densamente conectadas de pixeis, que podem corresponder a estruturas relevantes na imagem onde ocorre maior variação de intensidade. Como ilustrado pela Figura 36

Para cada uma das distribuições que vamos analisar podemos realizar esse mesmo processo de forma análoga, calculando a média, entropia e energia. Dessa forma para cada super pixel teremos duas assinaturas geradas por rede analisada, cada uma com as três medidas estatísticas e um prefixo indicando qual distribuição está sendo analisada, $[S, media, entropia, energia]$ onde o prefixo S indica que as medidas foram calculadas a partir da distribuição de força, $[C, media, entropia, energia]$ e C indica que foram calculadas por meio da distribuição do coeficiente de agrupamentos.

4.2.5 Gerando a representação final da imagem

Após a transformação de cada super pixel em uma rede complexa, aplicamos as medidas estatísticas de média, entropia e energia em cada distribuição. Essas medidas fornecem informa-

Figura 36 – Análise visual de como o coeficiente de agrupamento pode destacar regiões onde há maior variação de intensidade na imagem.

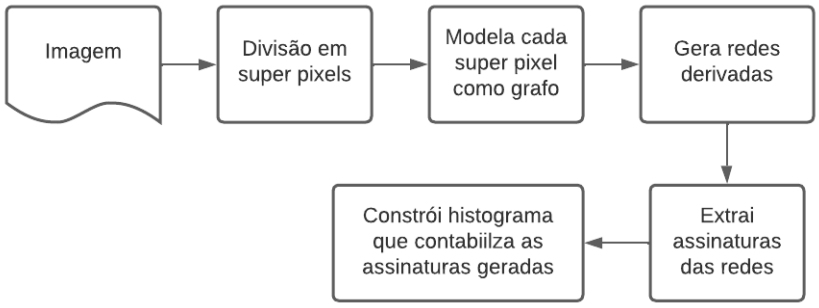
(a) representação da imagem original (b) representação com coeficiente de agrupamento

Fonte: Elaborada pelo autor.

ções importantes sobre o conteúdo de cada distribuição, permitindo que sejam sumarizadas em uma assinatura visual.

Em seguida, utilizamos essas assinaturas para construir um histograma que representa a frequência de ocorrência de cada assinatura. Esse histograma é uma representação compacta e eficiente da informação contida nas distribuições dos super pixeis da imagem original. Dessa forma, nosso método que pode ser resumido pela Figura 37, fornece uma descrição detalhada da textura da imagem, que pode ser utilizada em diversas aplicações.

Figura 37 – Metodologia passo a passo para a geração da representação final de uma imagem



Fonte: Elaborada pelo autor.

4.3 Resultados e Discussões

Nesta seção, apresentamos e discutimos os resultados obtidos ao aplicar o método proposto em sistemas de recuperação de imagens por conteúdo. Esses sistemas buscam automatizar a busca por imagens com base em suas características visuais.

O BoCS consiste em transformar cada super pixel em uma rede complexa e extrair estatísticas das distribuições resultantes. Foram feitos experimentos para a escolha dos melhores parâmetros e para comparar o método com técnicas baseadas nos paradigmas do BoVW. A implementação do código foi feita em Python e todos os experimentos foram feitos em um computador com processador Intel Core i5-8265U 1.6GHz, 16GB de RAM, e utilizando como sistema operacional o Ubuntu 18.04.

Para determinar a performance dos métodos, utilizamos as métricas precisão até os primeiros 10 elementos (P@10) e *Mean Average Precision* (MAP), como *baseline* utilizamos o C-BoVW (Sivic; Zisserman, 2003) e o BoSS (CHINO *et al.*, 2018). Para o C-BoVW extraímos as características das regiões de interesse utilizando o SIFT com dicionario de 1000 palavras visuais, para o BoSS utilizamos as configurações padrões que foram definidas em (CHINO *et al.*, 2018).

Para medir o desempenho do método testamos o seu desempenho em três diferentes bases de dados. A primeira delas é a base MRIBalan, composta por 704 imagens de ressonâncias magnéticas obtidas pelo Hospital das Clínicas de Ribeirão Preto da USP (Balan *et al.*, 2005). A segunda base utilizada foi a Covid-19 radiography database que é composta por 219 imagens com casos positivos de COVID-19 , 1341 imagens normais and 1345 imagens de pneumonia viral (Chowdhury *et al.*, 2020), por último utilizamos a base de texturas University of Illinois Urbana Champaign (UIUC) (LAZEBNIK; SCHMID; PONCE, 2005) composta por 25 classes com 40 imagens em cada uma.

4.3.1 Definindo os parâmetros do método

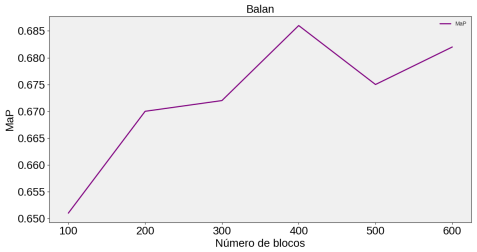
4.3.1.1 Definindo a quantidade de blocos

A escolha do número de super pixeis é uma questão importante no desenvolvimento do método proposto. Nesse sentido, é comum que essa definição seja realizada de forma empírica, buscando-se avaliar o desempenho do método para diferentes valores desse parâmetro.

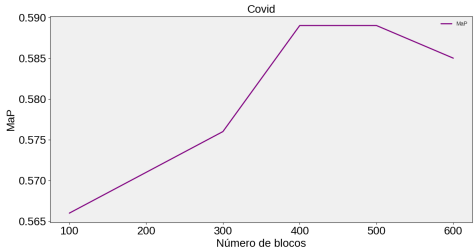
No caso do BoCS, variamos a quantidade de super pixeis em cada imagem em um intervalo de 100 a 600, a fim de encontrar o valor ideal para cada uma das bases de dados analisadas. Observamos que o pico de desempenho do método foi alcançado com valores entre 400 e 600 super pixeis, como pode ser visto na Figura 38.

Vale destacar que essa escolha do número de super pixeis ideal pode variar para diferentes bases de dados e aplicações, e que a determinação desse parâmetro pode ser considerada um desafio em si mesmo na análise de imagens por conteúdo. No entanto, a identificação de um intervalo inicial de valores que apresentam bons resultados, como o observado no presente estudo, pode ser útil para a aplicação do método em diferentes contextos.

Figura 38 – Análise da performance do método de acordo com a quantidade de super pixeis utilizada de acordo com a base dados.



(a) Variação da MAP de acordo com a quantidade de blocos utilizada para a base de dados Balan



(b) Variação da MAP de acordo com a quantidade de blocos utilizada para a base de dados Covid

(c) Variação da MAP de acordo com a quantidade de blocos utilizada para a base de dados Texturas

Fonte: Elaborada pelo autor.

4.3.1.2 Definindo o melhor raio

Ao modelar uma imagem como uma rede complexa, é essencial definir cuidadosamente os parâmetros utilizados na sua construção, pois isso afeta diretamente a qualidade e eficácia do método proposto. No caso da escolha do raio r que determina quais pixeis serão conectados na criação do super pixel como um grafo, é fundamental encontrar um valor ideal que permita uma boa representação da imagem. Para isso, realizamos experimentos variando o valor de r dentro do intervalo de 1 a 6 (sendo 1 a menor distância possível e 6 o raio médio dos super pixeis na base Balan quando usamos 400 blocos) além disso, utilizamos a maior distância entre dois pixeis no bloco para garantir a avaliação dos casos em que todos os pixeis estão conectados. Dessa forma, buscamos identificar o valor ótimo de r que proporciona uma representação precisa de cada super pixel como uma rede complexa.

Ao avaliar os resultados obtidos nas diferentes bases de dados utilizadas, observamos que, em geral, o desempenho do método é melhorado quanto maior for o valor de r. Entretanto, a escolha do melhor valor de r varia de acordo com a base de dados analisada. A Tabela 1 apresenta os resultados dos experimentos realizados, onde podemos verificar que, em quatro dos seis cenários analisados, utilizar a maior distância entre dois pixeis no bloco como valor de r é a melhor opção.

Tabela 1 – Métricas de avaliação do BoCS variando r

Raio	P@10(Balan)	P@10(Covid)	P@10(Textura)	MaP(Balan)	MaP(Covid)	MaP(Textura)
1	0,806	0,676	0,657	0,545	0,541	0,651
2	0,890	0,705	0,761	0,620	0,578	0,695
3	0,903	**0,765**	0,772	0,634	0,587	0,701
4	0,899	0,735	0,769	0,643	0,587	0,700
5	0,904	0,730	0,768	0,648	**0,589**	0,702
6	0,905	0,725	0,773	0,653	0,587	0,704
Max	**0,930**	0,708	**0,774**	**0,686**	0,582	**0,707**

4.3.1.3 Definindo os limiares

A definição do conjunto de limiares $T = \{t_1, t_2, ..., t_n\}$ é um passo crucial para o método proposto, pois a escolha do valor do limiar é crucial para obter uma rede complexa que capture adequadamente as características de textura. O trabalho realizado em (SCABINI *et al.*, 2019) apresenta uma abordagem automática para a sua determinação. O método leva em conta a variação de intensidade presente nas imagens. Arestas que conectam pixeis com valores de intensidade similares são predominantes na maior parte das imagens e portanto descartar arestas com um baixo peso iria destacar as conexões entre vértices(pixeis) com maior diferença de intensidade(destacando a textura nas imagens).

A Figura 39 mostra a distribuição de pesos das arestas nas bases de dados avaliadas durante esse trabalho. Como podemos ver as arestas com menor peso são predominantes em todas as bases. Como vimos anteriormente na Subseção 4.2.2 nossas redes complexas utilizam a diferença de intensidade dos pixeis para definir o peso das arestas, logo aquelas que tem menor peso são as arestas que conectam pixeis similares. Nosso objetivo com a limiarização é gerar novas redes onde possamos analisar a textura da imagem, logo ao descartar arestas com baixa variação de intensidade estamos destacando aquelas que tem uma grande variação e consequentemente destacando a textura dessa região. Portanto de acordo com (SCABINI *et al.*, 2019) podemos utilizar como limite inferior para o nosso *thresholding* $t_1 = argmax(P(w))$ ou seja, o peso mais frequente para as arestas na base de dados.

Ainda analisando a Figura 39 pode-se perceber que de forma geral, quanto maior o peso menor o número de arestas, isso significa que ao aplicar um limiar que corresponde a um peso alto o grau médio da rede iria diminuir. Redes com grau médio menor ou igual a um são consideradas pouco relevantes para análise de texturas, uma vez que possuem distribuições esparsas e muitos vértices não terão nenhuma conexão.

Portanto de acordo com (SCABINI *et al.*, 2019) para excluir da nossa análise essas redes pouco relevantes podemos encontrar o momento onde ao aplicar um limiar t_m teremos redes com grau médio menor ou igual a 1, ou seja no momento onde cada vértice tem em média apenas uma conexão. Considere $\gamma(w)$ como a função que nos dá o grau médio da rede ao se aplicar um determinado limiar w, o nosso limite superior pode ser definido como $t_m = argmax(\delta(\gamma(w), 1))$ de forma que δ é a função que retorna 1 quando $\gamma(w) = 1$ e 0 caso contrário, neste caso *argmax*

retorna o valor de w onde a rede tem grau médio 1.

Utilizando os limites que foram determinados de forma automática podemos dividi-los em m intervalos equidistantes, onde m pode ser definido pelo usuário. Neste trabalho utilizamos $m = 4$ gerando o conjunto $T = \{t_1, t_2, t_3, t_4\}$ de limiares.

4.3.2 Avaliando o impacto das funções de distância

Nesta seção, serão analisados os resultados obtidos em experimentos realizados com sistemas de recuperação de imagens por conteúdo, tendo como objetivo investigar o impacto de diferentes funções de distância na comparação entre as imagens. Avaliamos todas as funções de distância com base no MaP e P@10 em cada uma das bases de dados.

Considerando a Tabela 2 e a Tabela 3, é possível observar que as escolhas da função de distância em diferentes bases de dados tiveram um impacto significativo no desempenho do método BoCS. Com base nas métricas apresentadas, é possível identificar qual função de distância apresentou melhor desempenho para cada base de dados em particular.

Por exemplo, na base de dados Balan, a função de distância *Bhattacharyya* apresentou o melhor desempenho em termos de MAP e P@10. Já para a base de dados Covid, a melhor função de distância varia de acordo com a métrica, enquanto para a base de dados de textura, a função de distância *Jaccard* obteve os melhores resultados.

Tabela 2 – Avaliando o MaP para as diferentes funções de distâncias em todas as bases de dados

Base de dados	Euclidiana	City-block	Chebyshev	Hamming	Jaccard	Cosseno	Bhattacharyya
Balan	0.558	0.686	0.513	0.545	0.616	0.562	**0.715**
Covid	0.552	0.589	0.540	0.436	0.551	0.524	**0.598**
Textura	0.655	0.707	0.527	0.481	**0.725**	0.651	0.714

Tabela 3 – Avaliando o P@10 para as diferentes funções de distâncias em todas as bases de dados

Base de dados	Euclidiana	City-block	Chebyshev	Hamming	Jaccard	Cosseno	Bhattacharyya
Balan	0.807	0.930	0.969	0.704	0.787	0.854	**0.937**
Covid	0.694	**0.765**	0.636	0.420	0.617	0.698	0.722
Textura	0.735	0.774	0.603	0.497	**0.804**	0.726	0.781

Logo podemos observar que os resultados variaram significativamente entre as bases de dados e funções de distância avaliadas, indicando que não há uma função de distância única que seja a melhor para todas as situações. Portanto, é importante levar em consideração as características das bases de dados e dos conjuntos de imagens específicos ao selecionar a função de distância mais adequada para cada situação. Conclui-se, portanto, que a escolha adequada da função de distância é crucial para o sucesso do método proposto.

4.3.3 Comparações com outros métodos

Com o objetivo de avaliar o desempenho do método BoCS proposto, realizamos um conjunto de experimentos comparativos utilizando bases de dados de diferentes domínios: imagens de texturas e imagens médicas. A escolha dessas bases de dados foi motivada pelo fato de que, em cada uma dessas áreas, as características das imagens podem ser bastante distintas e, portanto, permitem uma avaliação mais abrangente do método proposto. Realizamos análises de desempenho em termos das métricas P@10 e MaP. Além disso, comparamos o desempenho do método proposto com os de outros métodos estabelecidos na literatura, a fim de avaliar a sua eficácia. Nesta seção, descreveremos os detalhes dos experimentos e resultados obtidos.

Com base nos experimentos descritos na seção anterior definimos os parâmetros do BoCS para cada uma das bases de dados, a Tabela 4 mostra com mais detalhes os parâmetros escolhidos.

Tabela 4 – Escolha dos parâmetros para cada uma das bases de imagens

Base de dados	Número de blocos	r	função de distância
Balan	400	Max	Bhattacharyya
Covid	400	3	Bhattacharyya
Textura	400	Max	Jaccard

Para a definição dos limiares de *thresholding* utilizamos o método automático de (SCABINI *et al.*, 2019), com $m = 4$, além disso as assinaturas visuais foram construídas sumarizando o conteúdo das distribuições de força das arestas e coeficiente de agrupamento dos vértices, com medidas como entropia, energia e média.

Ao comparar a efetividade de diferentes métodos de análise de imagens, é importante estabelecer um ponto de referência ou *baseline* para avaliar a performance dos novos métodos propostos. Neste trabalho, utilizamos como *baseline* um método do paradigma C-BoVW e outro do S-BoVW. Para o C-BoVW durante a fase de detecção de regiões de interesse utilizamos o descritor SIFT e criamos um dicionário de 1000 palavras visuais utilizando o algoritmo *K-means*, por fim utilizamos o histograma de palavras visuais como descritor de cada imagem. Para o paradigma S-BoVW, foi feita a comparação com o BoSS (CHINO *et al.*, 2018), método que até então tem a melhor performance dentro do paradigma. Para o BoSS utilizamos as configurações padrões que foram definidas em seu artigo de origem.

Na realização dos experimentos, utilizamos três diferentes bases de dados: Balan, Covid e Texturas UIUC. Na base de dados Balan, todas as imagens foram utilizadas como centros de consulta e busca, totalizando 704 consultas por conteúdo, onde cada imagem foi comparada com todas as outras imagens da base de dados. Já na base de dados Covid, com o objetivo de balancear o número de imagens por classe, selecionamos aleatoriamente um subconjunto de 219 imagens de cada classe, formando assim uma nova base de dados composta por 657 imagens, também com todas as imagens sendo utilizadas como centros de consulta e busca. Para a base de

dados de texturas UIUC, selecionamos aleatoriamente 6 classes, cada uma possui 40 imagens, portanto a base de dados resultante tem 240 imagens.

Pela Figura 40 podemos observar os resultados obtidos com o método BoCS ao compará-lo com outros métodos populares em recuperação de imagens. Utilizando a métrica P@10 para avaliar a eficácia da recuperação de imagens, o BoCS foi igual ou melhor que seus concorrentes em todas as três bases de dados avaliadas. Isso mostra que o BoCS é capaz de recuperar as imagens relevantes com alta precisão entre os primeiros 10 resultados.

Quando avaliamos a métrica MaP estamos olhando para o desempenho do método no geral, isto é, em todas as posições da sua busca. Com essa métrica o nosso método teve melhores resultados em duas das três bases avaliadas. Esses resultados indicam que o método proposto tem grande potencial para melhorar a recuperação de imagens e pode ser uma excelente opção em aplicações práticas.

A curva *precision-recall* (precisão e revocação) é uma métrica comumente utilizada para avaliar a qualidade do resultado da recuperação de imagens em CBIR. Comparar a curva do método proposto com as curvas dos outros métodos pode ajudar a avaliar se o método é competitivo em termos de precisão e revocação.

Com base nas curvas de *precision-recall* obtidas para os três métodos nas bases de dados analisadas, e apresentadas na Figura 41 podemos observar com mais cuidado as diferenças de desempenho. Na base de dados Balan, o método BoCS apresentou o melhor desempenho em geral. O método BoSS-T apresentou desempenho semelhante, porém com uma queda mais acentuada na precisão à medida que o *recall* aumenta. Já os métodos Boss-CT, BoVW e Boss tiveram um desempenho inferior em comparação com os outros dois métodos.

Na base de dados Covid, o BoSS-CT e BoSS se destacaram com um desempenho superior ao avaliar toda a curva do gráfico. Vale ressaltar que tanto o BoCS quanto o BoSS-T apresentaram desempenhos similares, com as maiores precisões durante a busca dos primeiros elementos, mostrando que são as melhores escolhas para situações onde poucas imagens serão retornadas na busca.

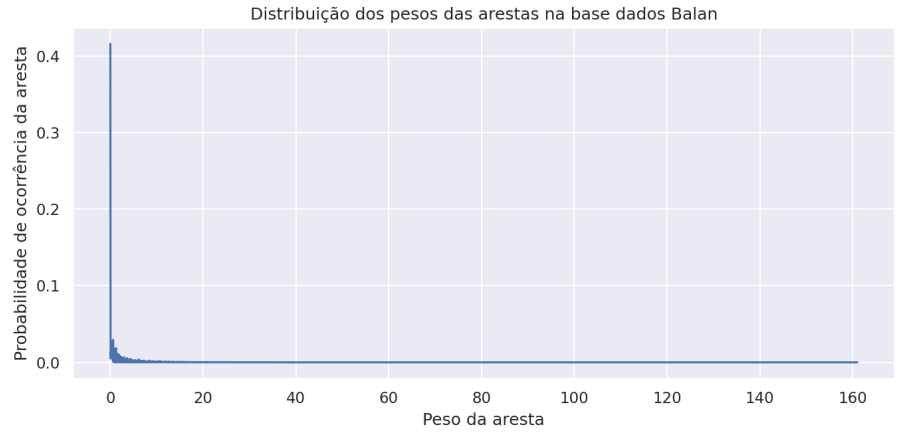
Na base de dados Textura, novamente o BoCS, mostrase o método com melhor desempenho nos primeiros elementos e teve desenho similar ao BoVW quando comparamos a curva geral. Enquanto os outros métodos tiveram quedas mais abruptas de precisão enquanto o *recall* aumenta.

4.4 Considerações Finais

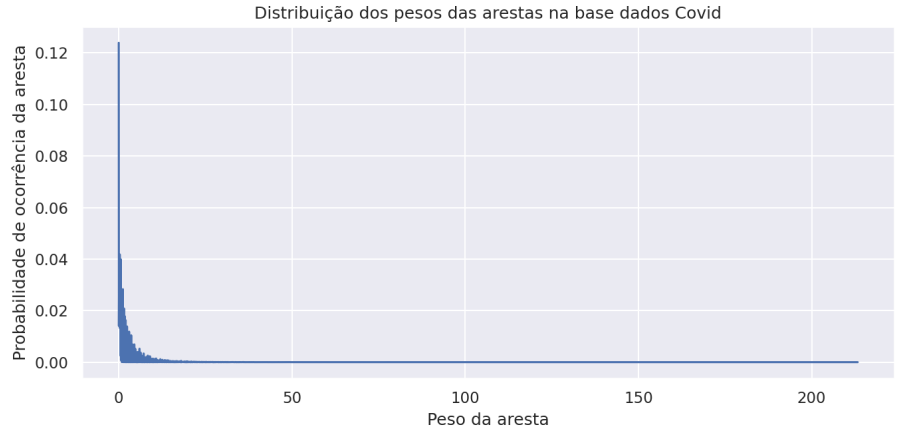
Este capítulo apresentou o método BoCS, que é a principal contribuição deste trabalho de Mestrado. O BoCS traz uma abordagem baseada em redes complexas, cujas assinaturas permitem a recuperação por conteúdo de modo mais preciso. O próximo capítulo irá apresentar

as conclusões e sugestões de trabalhos futuros.

Figura 39 – Distribuição de pesos das arestas para as bases de dados analisadas neste trabalho, podemos ver pelos gráficos que em todos os casos arestas com menor peso nas arestas são as mais frequentes nas bases de dados



(a) Distribuição da probabilidade de ocorrência dos pesos na base de dados Balan

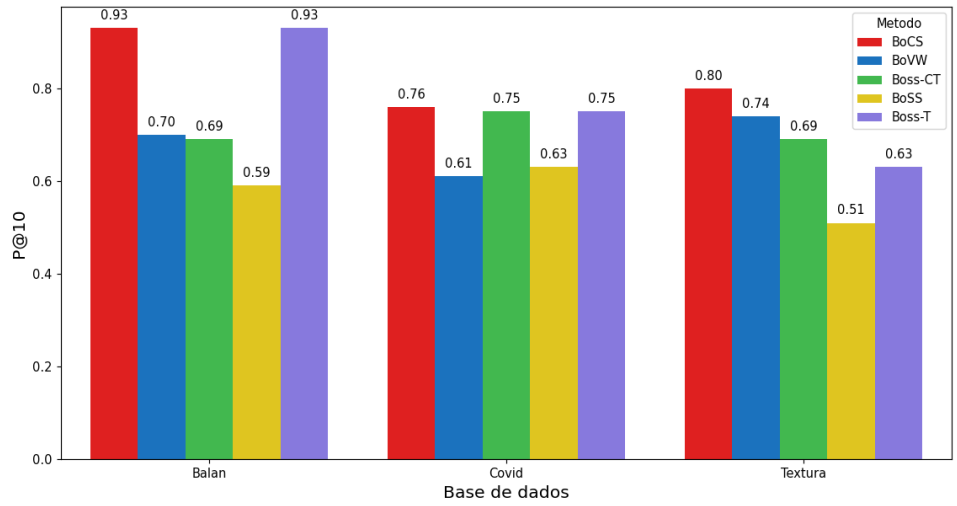


(b) Distribuição da probabilidade de ocorrência dos pesos na base de dados Covid

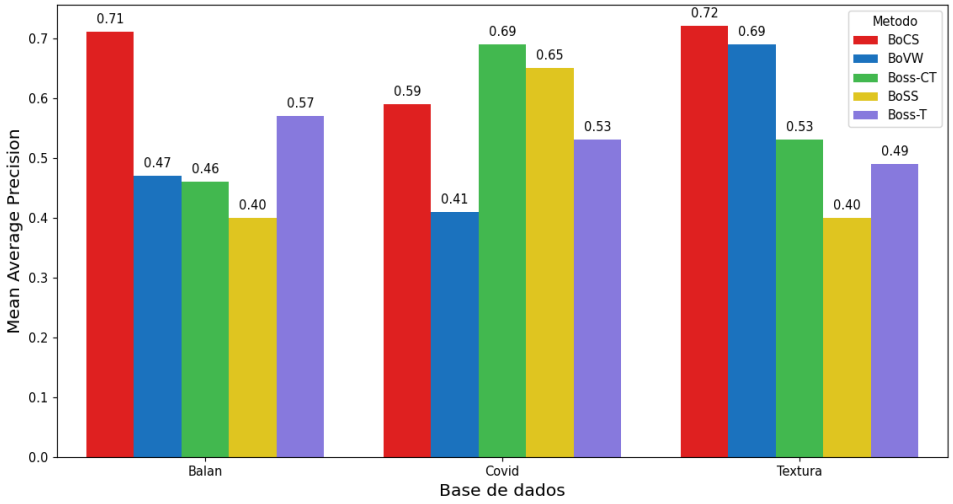
(c) Distribuição da probabilidade de ocorrência dos pesos na base de dados Textura

Fonte: Elaborada pelo autor.

Figura 40 – Resultados dos métodos por base de dados



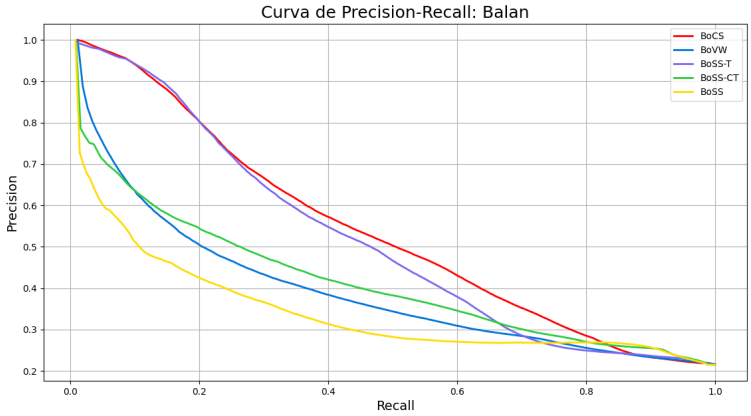
(a) Comparação dos métodos utilizando a métrica P@10



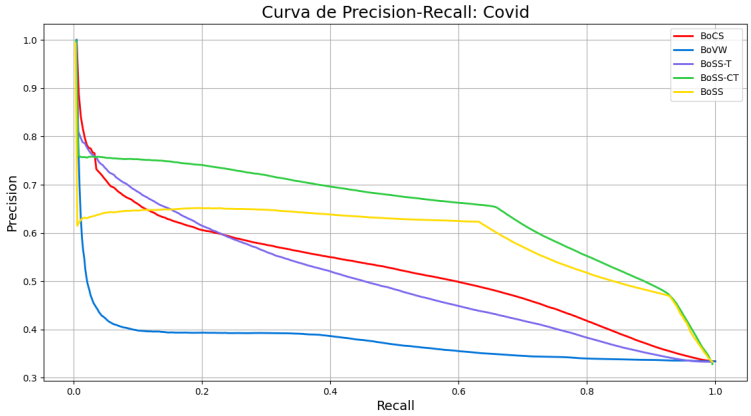
(b) Comparação dos métodos utilizando a métrica MaP

Fonte: Elaborada pelo autor.

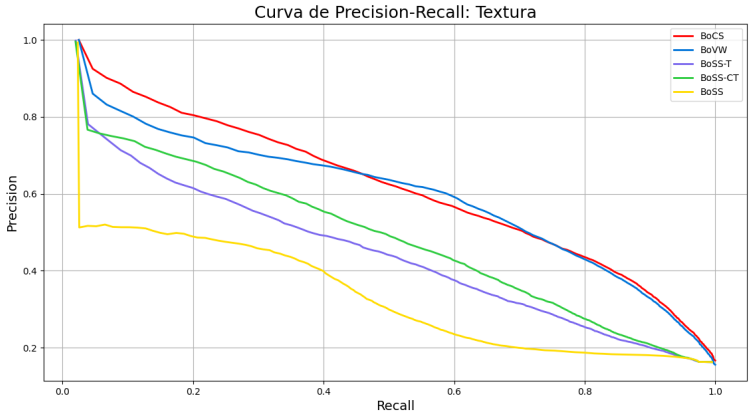
Figura 41 – Curva de *precision-recall* comparando o desempenho geral dos métodos



(a) Base de dados: Balan



(b) Base de dados: Covid



(c) Base de dados: Textura

Fonte: Elaborada pelo autor.

CAPÍTULO

5

CONCLUSÕES E TRABALHOS FUTUROS

5.1 Conclusões

Neste trabalho de mestrado, apresentamos o método BoCS, uma abordagem baseada em assinaturas visuais que utiliza redes complexas para extrair as características visuais das imagens. Realizamos seis cenários de comparação com outros métodos populares em recuperação de imagens, e em quatro desses cenários, o BoCS se mostrou superior de acordo com as métricas avaliadas. Os resultados obtidos comprovam a eficácia do método proposto em recuperar imagens relevantes com alta precisão. Além disso, a metodologia empregada na construção das assinaturas visuais e na utilização de redes complexas se mostrou promissora para a área de recuperação de imagens.

É importante destacar que o BoCS ainda apresenta limitações, como o grande número de parâmetros que pode impactar diretamente em sua performance. Devido ao escopo do trabalho de mestrado, não foram concentrados esforços na direção de mitigação do número de parâmetros para o método. Mas pode-se verificar que a abordagem por assinaturas visuais traz benefícios reais para o processamento de consultas por similaridade, utilizando técnicas de recuperação de imagens por conteúdo.

5.2 Futuras linhas de pesquisa

Sugestões para futuras pesquisas incluem a exploração de um conjunto diferente de propriedades dos grafos para a descrição de texturas, explorar a importância de cada assinatura visual para a representação da imagem, além de estender a análise do método para outras bases de dados. Por exemplo, a utilização de técnicas como *Graph Convolutional Networks* (GCN) para analisar as texturas de redes complexas modeladas a partir de imagens, já que as GCNs permitem a extração de informações de alto nível a partir das características da topologia do

grafo. Os GCN podem ser utilizados para explorar diferentes propriedades dos grafos, tais como a conectividade local, a estrutura global, entre outros aspectos, levando a uma análise mais abrangente das texturas presentes nas imagens (WU *et al.*, 2019).

As operações de convolução em grafos permitem que as características visuais sejam propagadas pelos nós da rede, levando em consideração a topologia da rede. Isso significa que as características de uma região da imagem podem ser compartilhadas com outras regiões que têm uma topologia semelhante na rede, permitindo que as GCNs capturem informações de contexto importantes.

Anda é possível explorar a importância dos vértices que estão sendo utilizados para construir as redes complexas como feito por (CANTERO *et al.*, 2020) no artigo os autores propõem uma extensão do algoritmo *PageRank* (PAGE *et al.*, 1999) para caracterizar a topologia das redes complexas e medir a importância das vértices, outra forma de medir a importância de vértices ou assinaturas visuais seria utilizando o método *Weighted Histogram* proposto por (PEDROSA, 2015) que define a importância das como assinaturas com base na região em que ela está, as com pouca relevância são aquelas que estão em regiões homogêneas, enquanto as mais relevantes estão em regiões não homogêneas como bordas ou com grande variações de textura.

A proposta do BoCS representa uma contribuição para a área de recuperação de imagens ao apresentar uma abordagem inovadora para a extração de características visuais e a utilização de redes complexas na descrição das imagens. Os resultados obtidos neste trabalho abrem caminho para novas pesquisas que possam expandir e aprimorar o método desenvolvido, contribuindo para o avanço da ciência na área de recuperação de imagens. Em síntese, o método BoCS apresentou resultados promissores, mas ainda há espaço para melhorias e novas pesquisas. Esperamos que este trabalho possa incentivar a comunidade acadêmica a desenvolver novas abordagens para a recuperação de imagens, utilizando a teoria dos grafos e técnicas de processamento de imagens.

REFERÊNCIAS

ACHANTA, R. *et al.* **SLIC superpixels**. [S.l.], 2010. Citado na página 61.

ALEMU, Y.; KOH, J.-b.; IKRAM, M.; KIM, D.-K. Image retrieval in multimedia databases: A survey. In: **2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing**. [S.l.: s.n.], 2009. p. 681–689. Citado na página 24.

ALKHAWLANI, M.; ELMOGY, M.; EL-BAKRY, H. Text-based, content-based, and semantic-based image retrievals: A survey. **International Journal of Computer and Information Technology**, v. 4, p. 58–66, 01 2015. Citado na página 23.

ALSMADI, M. K. Content-based image retrieval using color, shape and texture descriptors and features. **Arabian Journal for Science and Engineering**, Springer, v. 45, p. 3317–3330, 2020. Citado na página 27.

ALZU'BI, A.; AMIRA, A.; RAMZAN, N. Semantic content-based image retrieval: A comprehensive study. **Journal of Visual Communication and Image Representation**, Academic Press Inc., v. 32, p. 20–54, 8 2015. ISSN 10959076. Citado na página 32.

ARRUDA, H. F. d. **Multi-scale analysis of languages and knowledge through complex networks**. Tese (Doutorado) — Universidade de São Paulo, 2019. Citado na página 37.

AVALHAIS, L. P. S. **Transformação de espaços métricos otimizando a recuperação de imagens por conteúdo e avaliação por análise visual**. Dissertação (Mestrado) — Universidade de São Paulo, São Carlos, 2012. Citado na página 32.

BACKES, A. R.; CASANOVA, D.; BRUNO, O. M. Texture analysis and classification: A complex network-based approach. **Information Sciences**, v. 219, p. 168 – 180, 2013. ISSN 0020-0255. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0020025512004677>. Citado na página 50.

Balan, A. G. R.; Traina, A. J. M.; Traina, C.; Azevedo-Marques, P. M. Fractal analysis of image textures for indexing and retrieval by content. In: **18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)**. [S.l.: s.n.], 2005. p. 581–586. Citado na página 67.

BAY, H.; TUYTELAARS, T.; GOOL, L. V. SURF: Speeded up robust features. In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. Springer, Berlin, Heidelberg, 2006. v. 3951 LNCS, p. 404–417. ISBN 3540338322. ISSN 03029743. Disponível em: <https://link.springer.com/chapter/10.1007/11744023_32>. Citado na página 42.

CANTERO, S. V. A. B.; GONçALVES, D. N.; SCABINI, L. F. dos S.; GONçALVES, W. N. Importance of vertices in complex networks applied to texture analysis. **IEEE Transactions on Cybernetics**, v. 50, n. 2, p. 777–786, 2020. Citado nas páginas 50 e 78.

CAZZOLATO, M. T. **Conquering knowledge from images: improving image mining with region-based analysis and associated information**. Tese (Doutorado) — Universidade de São Paulo, São Carlos, 2019. Citado na página 29.

CHINO, D. Y.; SCABORA, L. C.; TRAINA, C.; TRAINA, A. J. BoSS: Image retrieval using bag-of-superpixels signatures. In: **Proceedings of the ACM Symposium on Applied Computing**. Pau: Association for Computing Machinery, 2018. p. 309–312. ISBN 9781450351911. Citado nas páginas 55, 67 e 71.

Chowdhury, M. E. H.; Rahman, T.; Khandakar, A.; Mazhar, R.; Kadir, M. A.; Mahbub, Z. B.; Islam, K. R.; Khan, M. S.; Iqbal, A.; Emadi, N. A.; Reaz, M. B. I.; Islam, M. T. Can ai help in screening viral and covid-19 pneumonia? **IEEE Access**, v. 8, p. 132665–132676, 2020. Citado na página 67.

CONNERS, R. W.; HARLOW, C. A. A theoretical comparison of texture algorithms. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, PAMI-2, n. 3, p. 204–222, 1980. ISSN 01628828. Citado na página 30.

COSTA, L. da F.; RODRIGUES, F.; TRAVIESO, G.; BOAS, P. V. Characterization of complex networks: A survey of measurements. **Advances in Physics**, v. 56, p. 167–242, 01 2007. Citado nas páginas 34, 35, 36, 38 e 39.

DUBEY, S. R. A decade survey of content based image retrieval using deep learning. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 32, n. 5, p. 2687–2704, 2022. Citado nas páginas 28 e 29.

FREEMAN, L. C. Centrality in social networks: Conceptual clarification. **Social networks**, Elsevier, v. 1, n. 3, p. 215–239, 1978. Citado na página 37.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing**. 2nd. ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2001. ISBN 0201180758. Citado na página 30.

HARALICK, R. M.; DINSTEIN, I.; SHANMUGAM, K. Textural Features for Image Classification. **IEEE Transactions on Systems, Man and Cybernetics**, SMC-3, n. 6, p. 610–621, 1973. ISSN 21682909. Citado na página 30.

JEGOU, H.; HARZALLAH, H.; SCHMID, C. A contextual dissimilarity measure for accurate and efficient image search. In: **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. Minneapolis: IEEE, 2007. ISBN 1424411807. ISSN 10636919. Citado na página 44.

JIANG, B.; LIU, D.; LI, Y.; ZHA, Z. A survey on content-based image retrieval: From handcrafted features to deep learning. **IEEE Access**, v. 9, p. 37883–37908, 2021. Citado na página 28.

JIANG, Y. G.; NGO, C. W.; YANG, J. Towards optimal bag-of-features for object categorization and semantic video retrieval. In: **Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR 2007**. Amsterdam: Association for Computing Machinery, 2007. p. 494–501. ISBN 1595937331. Citado na página 44.

JR, C.; TRAINA, A.; FALOUTSOS, C. Fast feature selection using fractal dimension - ten years later. **JIDM**, v. 1, p. 17–20, 01 2010. Citado na página 55.

KAILATH, T. The divergence and bhattacharyya distance measures in signal selection. **IEEE Transactions on Communication Technology**, v. 15, n. 1, p. 52–60, 1967. Citado na página 32.

LAZEBNIK, S.; SCHMID, C.; PONCE, J. A sparse texture representation using local affine regions. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 27, n. 8, p. 1265–1278, 2005. Citado na página 67.

LIMA, G. V. de; SAITO, P. T.; LOPES, F. M.; BUGATTI, P. H. Classification of texture based on Bag-of-Visual-Words through complex networks. **Expert Systems with Applications**, Elsevier Ltd, v. 133, p. 215–224, 11 2019. ISSN 09574174. Citado nas páginas 39, 56 e 57.

LOWE, D. G. Object recognition from local scale-invariant features. In: **Proceedings of the IEEE International Conference on Computer Vision**. Washington: IEEE, 1999. v. 2, p. 1150–1157. Citado na página 42.

____. Distinctive image features from scale-invariant keypoints. **International Journal of Computer Vision**, Springer, v. 60, n. 2, p. 91–110, 11 2004. ISSN 09205691. Disponível em: <https://link.springer.com/article/10.1023/B:VISI.0000029664.99615.94>. Citado nas páginas 42 e 43.

MIKOLAJCZYK, K.; SCHMID, C. Scale and affine invariant interest point detectors. **International Journal of Computer Vision**, v. 60, p. 63–86, 10 2004. Citado na página 42.

MUKHERJEE, D.; WU, Q. M. J.; WANG, G. A comparative experimental study of image feature detectors and descriptors. **Machine Vision and Applications**, Springer Verlag, v. 26, n. 4, p. 443–466, 5 2015. ISSN 14321769. Citado na página 41.

NEWMAN, M. E. J. The structure and function of complex networks. **SIAM review**, SIAM, v. 45, n. 2, p. 167–256, 2003. Citado nas páginas 34 e 39.

OJALA, T.; PIETIKÄINEN, M.; HARWOOD, D. A comparative study of texture measures with classification based on feature distributions. **Pattern Recognition**, Elsevier Ltd, v. 29, n. 1, p. 51–59, 1 1996. ISSN 00313203. Citado na página 31.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. The pagerank citation ranking: Bringing order to the web. **Stanford InfoLab**, v. 16, n. 1, p. 1–17, 1999. Citado nas páginas 50 e 78.

PEDROSA, G. V. **Caracterização e recuperação de imagens usando dicionários visuais semanticamente enriquecidos**. Tese (Doutorado) — Universidade de São Paulo, São Carlos, 2015. Citado nas páginas 24, 31, 42, 51, 52, 53, 54 e 78.

PHILBIN, J.; CHUM, O.; ISARD, M.; SIVIC, J.; ZISSERMAN, A. Lost in quantization: Improving particular object retrieval in large scale image databases. In: **26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR**. Anchorage: IEEE, 2008. ISBN 9781424422432. Citado na página 44.

SANTOS, J. M. dos. **Descritores de imagens baseados em assinatura textual**. Tese (Doutorado) — Universidade Federal do Amazonas, Manaus, 2016. Citado nas páginas 40, 45 e 46.

SANTOS, J. M. dos; MOURA, E. S. de; SILVA, A. S. da; CAVALCANTI, J. M. B.; TORRES, R. d. S.; VIDAL, M. L. A. A signature-based bag of visual words method for image indexing and search. **Pattern Recognition Letters**, North-Holland, v. 65, p. 1–7, 11 2015. ISSN 0167-8655. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167865515001956>. Citado nas páginas 45 e 64.

SCABINI, L. F.; CONDORI, R. H.; GONçALVES, W. N.; BRUNO, O. M. Multilayer complex network descriptors for color–texture characterization. **Information Sciences**, v. 491, p. 30–47, 2019. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025519301847>. Citado nas páginas 51, 52, 69 e 71.

SCABINI, L. F.; GONÇALVES, W. N.; CASTRO, A. A. Texture analysis by bag-of-visual-words of complex networks. In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. Montevideo: Springer Verlag, 2015. v. 9423, p. 485–492. ISBN 9783319257501. ISSN 16113349. Citado nas páginas 39 e 56.

SHAMNA, P.; GOVINDAN, V.; NAZEER, K. A. Content based medical image retrieval using topic and location model. **Journal of Biomedical Informatics**, Academic Press, v. 91, p. 103112, 3 2019. ISSN 1532-0464. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1532046419300309>. Citado na página 40.

SILVA, M. P. da. **Sistematização da percepção médica na construção de sistemas para recuperação de imagens por conteúdo**. Tese (Doutorado) — Universidade de São Paulo, São Carlos, 2014. Citado na página 32.

SINGH, S.; SINGH, M. P. An overview of image retrieval systems based on content and their limitations. **International Journal of Computer Science and Information Technology Research**, v. 8, n. 1, p. 68–75, 2020. Disponível em: <http://www.ijcsitre.org/volume8-issue1/ijcsitre-v8i1p9/>. Citado na página 28.

Sivic; Zisserman. Video Google: a text retrieval approach to object matching in videos. In: **Proceedings Ninth IEEE International Conference on Computer Vision**. IEEE, 2003. p. 1470–1477. ISBN 0-7695-1950-4. Disponível em: <http://ieeexplore.ieee.org/document/1238663/>. Citado nas páginas 40 e 67.

SOUZA, J. A. de. **Agrupamento de dados complexos para apoiar consultas por similaridade com tratamento de restrições**. Tese (Doutorado) — Universidade de São Paulo, São Carlos, 2019. Citado na página 33.

STRICKER, M. A.; ORENGO, M. Similarity of color images. In: NIBLACK, W.; JAIN, R. C. (Ed.). **Storage and Retrieval for Image and Video Databases III**. SPIE, 1995. v. 2420, p. 381. Disponível em: <http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.205308>. Citado na página 29.

STUTZ, D.; HERMANS, A.; LEIBE, B. Superpixels: An evaluation of the state-of-the-art. **Computer Vision and Image Understanding**, v. 166, p. 1 – 27, 2018. ISSN 1077-3142. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1077314217300589>. Citado na página 61.

SUN, W.; KISE, K. Similar manga retrieval using visual vocabulary based on regions of interest. In: **Proceedings of the International Conference on Document Analysis and Recognition, ICDAR**. Beijing: IEEE, 2011. p. 1075–1079. ISBN 9780769545202. ISSN 15205363. Citado na página 40.

SWAIN, M. J.; BALLARD, D. H. Color indexing. **International Journal of Computer Vision**, Kluwer Academic Publishers, v. 7, n. 1, p. 11–32, 1991. ISSN 15731405. Citado na página 29.

TAMURA, H.; MORI, S.; YAMAWAKI, T. Textural Features Corresponding to Visual Perception. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 8, n. 6, p. 460–473, 1978. ISSN 0018-9472. Disponível em: <http://ieeexplore.ieee.org/document/4309999/>. Citado na página 30.

VISHRAJ, R.; GUPTA, S.; SINGH, S. A comprehensive review of content-based image retrieval systems using deep learning and hand-crafted features in medical imaging: Research challenges and future directions. **Computers and Electrical Engineering**, v. 104, p. 108450, 2022. ISSN 0045-7906. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0045790622006656>. Citado na página 27.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. **Nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citado na página 36.

WU, Z.; PAN, S.; CHEN, F.; LONG, G.; ZHANG, C.; YU, P. S. Graph convolutional networks for image processing: A survey. **IEEE Transactions on Neural Networks and Learning Systems**, IEEE, v. 31, n. 10, p. 3204–3222, 2019. Citado na página 78.

ZAHROTUN, L. Comparison jaccard similarity, cosine similarity and combined both of the data clustering with shared nearest neighbor method. **Universitas Sriwijaya**, Vol 5 No 1 (2016), 2016. Citado na página 32.

ZHAO, H.; XU, Z.; HONG, P. Performance evaluation for three classes of textural coarseness. In: **Proceedings of the 2009 2nd International Congress on Image and Signal Processing, CISP'09**. Tianjin: IEEE, 2009. ISBN 9781424441310. Citado na página 30.


ICMC
USP
SÃO CARLOS