Caracterização e recuperação de imagens usando dicionários visuais semanticamente enriquecidos

Glauco Vitor Pedrosa

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Glauco Vitor Pedrosa

Caracterização e recuperação de imagens usando dicionários visuais semanticamente enriquecidos*

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Agma Juci Machado Traina

USP – São Carlos
Setembro de 2015

*Trabalho realizado com o apoio financeiro da FAPESP processo nº 2011/21460-4

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

P372c Pedrosa, Glauco Vitor
Caracterização e recuperação de imagens usando
dicionários visuais semanticamente enriquecidos /
Glauco Vitor Pedrosa; orientadora Agma Juci Machado
Traina. -- São Carlos, 2015.
109 p.

Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2015.

1. Extração de características. 2. Recuperação de
imagens. 3. Consultas por similaridade. 4.
Dicionários visuais. 5. Palavras visuais. I.
Traina, Agma Juci Machado, orient. II. Título.

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Glauco Vitor Pedrosa

Image characterization and retrieval using visual dictionaries semantically enriched *

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Agma Juci Machado Traina

USP – São Carlos
September 2015

*This work has been supported by FAPESP (Process Number 2011/21460-4)

Agradecimentos

Agradeço...

A Deus, por minha vida.

À minha família pela dedicação, confiança e apoio incondicional em todos os momentos.

À minha orientadora profa. Dra. Agma Juci Machado Traina, por me ensinar, incentivar, pelos conselhos e atenção que me auxiliaram no desenvolvimento desse trabalho, e por ser um exemplo a quem sempre seguirei.

Aos meus amigos do GBdI que se mostraram companheiros e, diretamente ou indiretamente, contribuíram para a realização deste trabalho.

Aos funcionários e professores do ICMC/USP.

À FAPESP (Processo Número 2011/21460-4), CAPES e CNPq pelo apoio financeiro à realização deste trabalho.

A análise automática da similaridade entre imagens depende fortemente de descritores que consigam caracterizar o conteúdo das imagens em dados compactos e discriminativos. Esses dados extraídos e representados em um vetor-de-características tem o objetivo de representar as imagens nos processos de mineração e análise para classificação e/ou recuperação. Neste trabalho foi explorado o uso de dicionários visuais e contexto para representar e recuperar as características locais das imagens utilizando formalismos estendidos com alto poder descritivo. Esta tese apresenta em destaque três novas propostas que contribuem competitivamente com outros trabalhos da literatura no avanço do estado-da-arte, desenvolvendo novas metodologias para a caracterização de imagens e para o processamento de consultas por similaridade.

A primeira proposta estende a modelagem *Bag-of-Visual-Words*, permitindo codificar a interação entre palavras-visuais e suas disposições espaciais na imagem. Para tal fim, três novas abordagens são apresentadas: (i) *Weighted Histogram* (WE); (ii) *Bunch-of-2-grams* e (iii) *Global Spatial Arrangement* (GSA). Cada uma dessas técnicas permitem extrair informações semanticamente complementares, que enriquecem a representação final das imagens descritas em palavras-visuais.

A segunda proposta apresenta um novo descritor, chamado de *Bag-of-Saliency-Points* (BoSP), que caracteriza e analisa a dissimilaridade de formas (silhuetas) de objetos explorando seus pontos de saliências. O descritor BoSP se apoia no uso de um *dicionário de curvaturas* e em *histogramas espaciais* para representar sucintamente as saliências de um objeto em um único vetor-de-características de tamanho fixo, permitindo recuperar formas usando funções de distâncias computacionalmente rápidas.

Por fim, a terceira proposta apresenta um novo modelo de consulta por similaridade, denominada *Similarity Based on Dominant Images* ($SimD_{Im}$), baseada no conceito de Imagens Dominantes, que é um conjunto que representa, de uma maneira mais diversificada e reduzida, toda a coleção de imagens da base de dados. Tal conceito permite dar mais eficiência quando se deseja analisar o contexto da coleção, que é o objetivo da proposta.

Os experimentos realizados mostram que os métodos propostos contribuem de maneira efetiva para caracterizar e quantificar a similaridade entre imagens por meio de abordagens estendidas baseadas em dicionários visuais e análise contextual, reduzindo a lacuna semântica existente entre a percepção humana e a descrição computacional.

Palavras-Chaves: Extração de características, Recuperação de imagens, Consultas por similaridade, Dicionários visuais, Palavras visuais.

Abstract

The automatic similarity analysis between images depends heavily on the use of descriptors that should be able to characterize the images' content in compact and discriminative features. These extracted features are represented by a feature-vector employed to represent the images in the process of mining and analysis for classification and/or retrieval. This work investigated the use of visual dictionaries and context to represent and retrieve the local image features using extended formalism with high descriptive power. This thesis presents three new proposals that contribute in advancing the state-of-the-art by developing new methodologies for characterizing images and for processing similarity queries by content.

The first proposal extends the Bag-of-Visual-Words model, by encoding the interaction between the visual words and their spatial arrangements in the image space. For this, three new techniques are presented: (i) Weighted Histogram (WE); (ii) Bunch-of-2-grams and (iii) Global Spatial Arrangement (GSA). These three techniques allow to extract additional semantically information that enrich the final image representation described in visual-words.

The second proposal introduces a new descriptor, called Bag-of-Saliency-Points (BoSP), which characterizes and analyzes the dissimilarity of shapes (silhouettes) exploring their salient point. The BoSP descriptor is based on using a dictionary of curvatures and spatial-histograms to represent succinctly the saliencies of a shape into a single fixed-length feature-vector, allowing to retrieve shapes using distance functions computationally fast.

Finally, the third proposal introduces a new similarity query model, called Similarity based on Dominant Images ($SimD_{Im}$), based on the concept of dominant images, which is a set of images representing the entire collection of images of the database in a more diversified and reduced manner. This concept allows to efficiently analyze the context of the entire collection, which is the final goal.

The experiments showed that the proposed methods effectively contributed to characterize and quantify the similarity between images using extended approaches based on visual dictionaries and contextual analysis, reducing the semantic gap between human perception and computational description.

Keywords: Feature Extraction, Image Retrieval, Similarity Search, Visual Dictionary, Visual Words.

Sumário

Lista de Figuras	xi
Lista de Tabelas	xv
Lista de Siglas e Acrônimos	xvii
Lista de Símbolos	xix
1 Introdução	1
1.1 Contextualização	1
1.2 Motivação	2
1.3 Definição do Problema	3
1.4 Objetivos e Contribuições	5
1.5 Organização do documento	6
2 Recuperação de Imagens Baseada em Conteúdo	9
2.1 Considerações Iniciais	9
2.2 Arquitetura de um Sistema CBIR	10
2.2.1 Estruturas de Indexação	10
2.2.2 <i>Gap</i> Semântico e a Realimentação de Relevância	11
2.3 Análise da Similaridade	12
2.3.1 Propriedades Métricas	12
2.3.2 Funções de Distância	13
2.3.3 Tipos de Consultas por Similaridade	14
2.4 Avaliação do Desempenho de um Sistema CBIR	15
2.4.1 Curva de Precisão vs. Revocação	15
2.4.2 Precisão Média	16
2.5 Descritores de Imagens	17
2.5.1 Descritores Globais	17
2.5.2 Descritores Locais	18
2.6 Considerações Finais	23
3 Descrevendo Imagens em Palavras-Visuais	25
3.1 Considerações Iniciais	25
3.2 A Abordagem <i>Bag-of-Visual-Words</i> (BoVW)	26
3.2.1 Metodologia	27
3.2.2 Construção do Dicionário Visual	29
3.2.3 Fase de Codificação	31
3.2.4 Fase de Sumarização	33
3.3 Avaliações Experimentais	34

3.3.1	Bases Avaliadas	34
3.3.2	Abordagens Avaliadas	35
3.3.3	Resultados Obtidos	36
3.4	Considerações Finais	39
4	Abordagens Estendidas para o Modelo <i>Bag-of-Visual-Words</i>	41
4.1	Considerações Iniciais	41
4.2	Trabalhos Correlatos	43
4.2.1	Codificação Espacial Baseada em Distância	43
4.2.2	Codificação Espacial Baseada em Regiões	46
4.3	Métodos Propostos: Agregando a Informação Espacial	47
4.3.1	Definição Formal do Problema	48
4.3.2	Distribuição Ponderada de Frequências	49
4.3.3	Co-ocorrências Baseada em Bag-of-Visual-Phrases	51
4.3.4	Arranjo Espacial Global	54
4.4	Avaliações Experimentais	58
4.4.1	Resultados Obtidos	59
4.5	Considerações Finais	61
5	Descrição de Formas em <i>Bag-of-Saliency-Points</i>	63
5.1	Considerações Iniciais	63
5.2	Contextualização	64
5.2.1	Desafios e Taxonomia	64
5.2.2	Motivação	66
5.3	Método Proposto	67
5.3.1	Detecção das Saliências	67
5.3.2	Representação das Saliências	69
5.3.3	Geração do Dicionário de Curvaturas	70
5.3.4	Descrição das Saliências em Palavras-Visuais	71
5.4	Usando Características Globais	73
5.5	Avaliações Experimentais	74
5.5.1	Bases de Imagens	74
5.5.2	Impacto dos Parâmetros	75
5.5.3	Resultados Obtidos	75
5.6	Considerações Finais	79
6	Consulta por Similaridade Apoiada por Imagens Dominantes	81
6.1	Considerações Iniciais	81
6.2	Contextualização	82
6.3	Método Proposto: Similaridade Baseada em Imagens Dominantes	84
6.3.1	Definição Formal do Problema	85
6.3.2	Definição de Imagens Dominadas e Dominantes	85
6.3.3	Etapa de Mineração	87
6.3.4	Etapa de Casamento	88
6.4	Avaliações Experimentais	90
6.5	Considerações Finais	93

7	Conclusão	95
7.1	Considerações Finais	95
7.2	Contribuições	96
7.3	Trabalhos Futuros	97
7.4	Publicações Geradas	98
	Referências Bibliográficas	101

Lista de Figuras

1.1	Abordagem clássica da metodologia <i>Bag-of-Visual-Words</i>	4
2.1	Metodologia de um sistema de recuperação de imagem por conteúdo. . . .	11
2.2	Conjuntos de imagens avaliadas na Recuperação de Imagens baseada em Conteúdo.	16
2.3	Metodologia para detectar as regiões-de-interesse de uma imagem: grade-regular ou pontos-de-interesse.	20
2.4	Pontos-de-interesse detectados pelas técnicas: (a) Harris , (b) Diferença de Gaussianas (DoG), (c) SUSAN.	22
2.5	O descritor SIFT é criado computando-se a magnitude do gradiente e a orientação da região em volta do ponto de interesse. Os gradientes são ponderados por uma janela Gaussiana, indicada pelo círculo. As amos- tras são acumuladas em histogramas de orientação (8 direções) para cada sub-região, que no exemplo mostrado à direita tem tamanho 4x4. O tama- nho das setas corresponde à soma das magnitudes dos gradientes próximos àquela direção. Imagem traduzida de (Lowe, 2004).	23
3.1	Metodologia clássica para representar uma imagem em <i>Bag-of-Visual-Words</i> . . .	27
3.2	Imagem de radiografia dividida em regiões (<i>patches</i>) de tamanho 9x9 pixels. Metodologia utilizada no trabalho de (Avni et al., 2011) para detectar as regiões-de-interesse.	29
3.3	Processo utilizado para a geração de um dicionário de palavras-visuais: dado um subconjunto de imagens de treinamento, suas características são extraídas e descritas. As palavras-visuais que irão compor o vocabulário do dicionário são definidas como sendo os centróides dos grupos a partir de um agrupamento realizado nesse espaço de características.	30
3.4	Qual palavra-visual será atribuída ao triângulo amarelo?	32
3.5	Exemplo da aplicação das técnicas de <i>pooling</i> (<i>avg</i> , <i>max</i> e <i>sum</i>) em uma imagem com 5 pontos e 3 palavras-visuais usando: (a) <i>soft assignment</i> e (b) <i>hard assignment</i>	33
3.6	Alguns exemplos de imagens de cada base utilizada nos experimentos. . . .	36
3.7	Resultado da aplicação dos detectores de pontos-de-interesse utilizados nos experimentos.	37
3.8	Valores de mAP (mean Average Precision) obtidos usando diferentes: ta- manho de dicionários, detectores de pontos-de-interesse e técnicas de codi- ficação.	38
4.1	Linha-do-tempo dos trabalhos correlatos.	44

4.2	A estratégia da técnica <i>Spatial Pyramid</i> (SP) é dividir a imagem em células hierárquicas e computar uma BoVW para cada região. Essa abordagem além de não ser invariante à rotação, gera um vetor com alta dimensionalidade.	47
4.3	Exemplo mostrando a proposta de ponderação de pontos. Quanto mais escuro, menos importante é o ponto e menor é a contribuição da palavra-visual associada a este ponto na representação final da imagem.	51
4.4	Metodologia <i>Bag-of-Visual-Words</i> e <i>Bag-of-Visual-Phrases</i>	52
4.5	Exemplo mostrando a construção do modelo <i>Bag-of-2-grams</i>	53
4.6	Um quadrante é definido utilizando cada ponto p_i detectado na imagem, então incrementa-se a ocorrência das palavras-visuais em cada região desse quadrante.	54
4.7	Exemplo da estratégia utilizada para codificar o arranjo espacial das palavras visuais. Os círculos vermelhos são os pontos detectados com a sua palavra-visual associada. Coloca-se a origem de um quadrante em cada ponto detectado e, então, conta-se, de maneira incremental, o número de palavras-visuais que aparecem em cada região do quadrante. Os valores finais do contador são exibidas na imagem inferior direita.	55
4.8	O gradiente sempre aponta para a região onde ocorre a maior variação de intensidade.	56
4.9	A direção do gradiente é usada para definir cada região do quadrante. Essa estratégia torna a representação invariante à rotação.	57
4.10	A representação final proposta agrupa cada região do quadrante em relação às suas posições: <i>top</i> , <i>down</i> , <i>left</i> e <i>right</i>	57
4.11	Valores de mAP (mean Average Precision) obtidos usando diferentes valores de r , em que μ e σ representam a média e o desvio padrão das distância entre os pontos detectados na imagem.	60
4.12	Valores de mAP (<i>mean Average Precision</i>) obtidos utilizando cada descritor. Os valores de <i>mAP1</i> se referem às bases originais e <i>mAP2</i> se referem às bases estendidas com imagens rotacionadas.	60
5.1	Formas de objetos diferentes porém com alta similaridade visual.	65
5.2	Classificação das técnicas de descrição e representação de formas.	65
5.3	Pontos de saliências de algumas formas.	66
5.4	Esquema do descritor BoSP.	68
5.5	(a) Uma forma com suas saliências; (b) Representação gráfica dos valores de curvatura das saliências usando diferentes valores de t na equação da curvatura (Eq. 5.1). O valor de m representa a quantidade de pontos do contorno.	70
5.6	Para codificar o arranjo espacial das palavras visuais, o espaço da forma é dividido em zonas de acordo com a distância a partir do centróide e então calcula-se um histograma em cada zona. Neste exemplo, utilizou-se três zonas e distância Euclidiana.	72
5.7	Mínimo retângulo delimitador e <i>Convex Hull</i> de uma forma.	73
5.8	A forma A é visualmente mais semelhante a forma B do que a C. No entanto, a forma B tem mais pontos salientes do que C, o que pode penalizar negativamente a similaridade. Duas características globais (excentricidade e solidez) são adicionadas para reduzir essa diferença.	74

5.9	mAP (<i>mean Average Precision</i>) valores obtidos usando tamanhos de dicionários diferentes e diferentes quantidades de zonas de distribuição de distância. (a) MPEG-7 (b) Kimia-216 (c) Kimia-99. Para $Z = 1$ considera-se apenas o Histograma Global.	76
5.10	Uma amostra do <i>ranking</i> gerado pela recuperação usando os descritores avaliados.	78
5.11	Curvas de PR para as bases: (a) Kimia-216 (b) Kimia-99.	79
5.12	Precisão do descritor BoSP+GF para a avaliação humana (qualitativo) e para a classificação definida da base (quantitativo).	79
6.1	Exemplo mostrando um subconjunto formado por imagens não-semelhantes, dada uma distância máxima de similaridade.	86
6.2	A imagem i é dita imagem dominada, pois a distância de i para j é menor que i para a imagem q , além disso a distância de j para q é maior que a distância de j para i	87
6.3	A figura (b) mostra os passos para a geração do conjunto de Imagens Dominantes do exemplo da figura (a). Dado um <i>ranking</i> inicial, a imagem mais próxima img_1 à imagem consulta img_q é inserida no conjunto (circulo amarelo), verifica-se quais imagens são dominadas por ela (círculos cinzas). Repete-se tal processo para outras imagens não dominadas e não inseridas no conjunto de Imagens Dominantes.	89
6.4	Valores de mAP (<i>mean Average Precision</i>) obtidos utilizando cada técnica de ponderação considerando diferentes funções de distâncias.	92

Lista de Tabelas

2.1	Alguns dos descritores mais utilizados em sistemas CBIR, em suas configurações tradicionais.	18
3.1	mAP valores obtidos usando diferentes funções de distâncias.	39
4.1	Tempo e espaço utilizado pelo descritores avaliados.	59
4.2	Ganho obtido em cada descritor avaliado quando comparado à técnica tradicional BoVW baseada em histogramas de palavras-visuais. Os valores de <i>mAP1</i> se referem às bases originais e <i>mAP2</i> se referem às bases estendidas com imagens rotacionadas.	61
5.1	Taxa <i>Bulls-Eye Score</i> e o tempo médio gasto na análise da similaridade para a base MPEG-7.	77
6.1	Ganhos obtidos em cada técnica quando comparada ao método tradicional de similaridade baseada em análise par-a-par sem nenhuma ponderação contextual (<i>baseline</i>).	93
6.2	Tempo médio exigido em cada técnica para extrair informações (<i>mining</i>) e realizar a análise da similaridade (<i>matching</i>).	94

Lista de Siglas e Acrônimos

Bo2g	<i>Bunch-of-2-grams</i>
BoSP	<i>Bag-of-Saliency-Points</i>
BoSP+GF	<i>Bag-of-Saliency-Points + Global Features</i>
BoVG	<i>Bag-of-Visual-Graphs</i>
BoVW	<i>Bag-of-Visual-Words</i>
BoW	<i>Bag-of-Words</i>
CBIR	<i>Content Based Image Retrieval</i>
DF	<i>Document Frequency</i>
DoG	<i>Diferença-de-Gaussianas</i>
DP	<i>Dynamic Programming</i>
GSA	<i>Global Spatial Arrangement</i>
IDF	<i>Inverse Document Frequency</i>
MAM	<i>Metric Access Methods</i>
<i>mAP</i>	<i>mean Average Precision</i>
MI	<i>Momentos Invariantes</i>
MS Fractal	<i>Multi-scale Fractal Dimension</i>
PIWAH	<i>Pairwise Identical Visual Words</i>
QBE	<i>Query by Example</i>
QBIC	<i>Query By Image Content</i>
QPM	<i>Query Point Movement</i>
SIFT	<i>Scale Invariant Feature Transform</i>
<i>SimD_{Im}</i>	<i>Similarity Based on Dominant Images</i>
SP	<i>Spatial Pyramid</i>
SSD	<i>Shape Saliency Descriptor</i>
TF	<i>Term Frequency</i>
TF-IDF	<i>Term Frequency-Inverse Document Frequency</i>

VLAD	<i>Vector of Locally Aggregated Descriptors</i>
WH	<i>Weighted Histogram</i>
WSA	<i>Words Spatial Arrangement</i>

Lista de Símbolos

$\mathcal{O}$	conjunto de pontos de um espaço vetorial n -dimensional
d	função de distância
L_p	distância de Minkowski
L'_p	distância de Minkowski ponderada
χ^2	distância χ^2
IH	distância Intersecção de Histogramas
$d_{Canberra}$	distância Canberra
$d_{Cosseno}$	distância Cosseno
R_q	<i>Range query</i>
kNNq	<i>k-Nearest Neighbor query</i>
$\mathbf{R}$	conjunto de imagens relevantes
$\mathbf{A}$	conjunto de imagens recuperadas
$\mathbf{Ra}$	conjunto de imagens relevantes recuperadas
$\mathbf{Pr}$	Precisão
$\mathbf{Re}$	Revocação
I	imagem
x,y	coordenadas x e y em uma imagem
m	magnitude
L	orientação
$P = \{p_1, p_2, \dots, p_N\}$	conjunto não-ordenado de N pontos-de-interesse detectados
N	quantidade de pontos-de-interesse detectados
D	dimensionalidade vetor-de-características
$C = \{c_1, \dots, c_K\}$	dicionários de palavras-visuais
K	tamanho do dicionário de palavras
α_j^i	função de ativação
$g : \mathbb{R}^{N \times K} \rightarrow \mathbb{R}^K$	função de pooling

H	histograma
w	função de ponderação
$\phi_{vizinho}$	conjunto dos vizinhos
$\phi_{similar}$	conjunto dos vizinhos similares
2-grams	conjunto de 2-grams
$f : \mathbb{R}^D \rightarrow \mathbb{R}^K$	função de codificação
$\mathcal{Q} = \{Q1, Q2, Q3, Q4\}$	quadrante
$G(p_i) = \{G_x(p_i), G_y(p_i)\}$	gradiente de um ponto p_i
$\tilde{B}$	conjunto dos pontos do contorno de uma forma
Ψ_t	função de curvatura discreta
u	filtro anisotrópico
s_i	ponto de saliência
Z	quantidade de zonas de distâncias
E	excentricidade
S	solidez
img_q	imagem de consulta
R_q	<i>ranking</i> de imagens
δ_{Sim}	função de similaridade
D_{Im}	conjunto de imagens dominantes
ϵ	função que extrai um vetor-de-características
$\delta_{SimD_{Im}}$	função de similaridade $SimD_{Im}$

Introdução

1.1 Contextualização

É inegável a importância das imagens no mundo moderno. Além de fontes de informação e transmissão de conhecimento, o uso de imagens digitais também é imprescindível em algumas áreas específicas tais como medicina, publicidade, gestão de produção, controle de segurança, e até na área de entretenimento. Por esse motivo não é de se estranhar o poder gigantesco de influência que elas exercem sobre o ser humano. Com o rápido e contínuo avanço das tecnologias que auxiliam na aquisição e no armazenamento de imagens, está havendo um aumento significativo de conteúdo visual gerado e disponibilizado. Por exemplo, um dado estatístico disponibilizado por uma rede social de compartilhamento de fotos, *Instagram*¹, mostra que a quantidade de fotos publicadas por todos os seus usuários supera a marca de 60 milhões por dia. Como consequência, nos dias atuais há uma grande necessidade de criação de sistemas de busca por informação, especificamente por dados complexos e não estruturados, tais como imagens.

Dependendo do formato da busca, a recuperação de imagens é dividida em duas categorias: baseada em texto e baseada em conteúdo. A primeira categoria é associada a palavras-chaves armazenadas com cada imagem, nesse caso o usuário normalmente só precisa digitar as palavras-chaves para realizar a busca. No entanto, por vezes, pode ser difícil e trabalhoso descrever o conteúdo de uma imagem em um pequeno conjunto de palavras-chaves, o que motiva a recuperação baseada em conteúdo, que visa mapear de uma maneira automática a semelhança semântica do conteúdo entre as imagens na busca pela informação. Essa abordagem é realizada por Sistemas de Recuperação de Imagens

¹disponível em: <http://instagram.com/press/>

por Conteúdo (CBIR - *Content Based Image Retrieval*), que são sistemas projetados para fazer a recuperação de imagens utilizando suas características visuais intrínsecas. Basicamente, um sistema CBIR possui a função de retornar as imagens de uma base de dados mais similares à uma dada imagem de consulta, considerando suas propriedades visuais.

Para diversas áreas do conhecimento humano os sistemas CBIR são ferramentas muito valiosas. Na área médica, por exemplo, esses sistemas auxiliam no diagnóstico médico, pois permitem a recuperação de imagens/exames mais semelhantes àqueles que o especialista médico está analisando (Hwang et al., 2012). Porém, ao tentar recuperar uma imagem usando um sistema CBIR, o especialista médico pode se deparar com o problema da quantidade excessiva de imagens irrelevantes recuperadas, seja por dificuldade da escolha de uma imagem adequada para consulta ou pela falta de descritores apropriados que consigam incorporar a subjetividade da percepção visual humana.

O grande desafio é que o problema de analisar a similaridade entre imagens por meio de um sistema computacional não é uma tarefa fácil, pois requer um descritor que capture a essência da imagem. Esse descritor deve extrair das imagens um número adequado de características para representá-la, de tal maneira que essas características extraídas sejam suficientemente discriminatórias para diferenciá-la de outras imagens. Essa situação deve ser levada em consideração para se obter um sistema CBIR com desempenho similar à percepção humana. Baseando-se nessas necessidades, diversas pesquisas estão sendo realizadas com o objetivo de desenvolver métodos eficientes e eficazes, que aumentem o poder discriminativo e diminuam o tempo de resposta de uma consulta em um sistema CBIR.

1.2 Motivação

A representação das características visuais de uma imagem é uma tarefa importante no processo de análise, classificação e recuperação. A extração dos atributos visuais sintetiza a imagem por meio de um *vetor-de-características*, que passa a representá-la nos processos de mineração e recuperação por conteúdo. Esse vetor-de-características deve ser combinado com uma *função de distância*, cujo objetivo é calcular o grau de similaridade entre dois vetores, retornando um valor que quantifica o quão similar eles são. Desse modo, segundo (Torres e Falcão, 2006), um descritor é um par (vetor-de-características, função de distância) utilizado para representar e recuperar imagens.

Descritores podem ser *globais* ou *locais*, dependendo de como analisam o conteúdo visual da imagem. A questão é que o desenvolvimento de um descritor é uma tarefa muito desafiadora, pois existem muitos fatores que devem ser considerados na descrição de uma imagem, tais como: resolução, as variações de iluminação, oclusões parciais, deformações e diferentes posições do sensor durante a captura da imagem. Os descritores globais são

mais populares na literatura, mas são conhecidos por serem limitados a esses fatores, pois descrevem uma imagem de uma forma holística, e para algumas aplicações essa abordagem de descrição pode não ser adequada, por exemplo, quando se deseja representar apenas um pequeno objeto contido em uma imagem em grande escala. Por outro lado, os descritores locais, como o descritor SIFT (*Scale Invariant Feature Transform*), permitem descrever várias regiões ao redor de pontos-de-interesse dentro das imagens, propiciando um grande poder descritivo, uma vez que é possível caracterizar apenas regiões específicas da imagem que sejam de interesse do usuário, excluindo informações irrelevantes/desnecessárias. Porém, a desvantagem do uso de descritores locais está no fato de que a cardinalidade do conjunto de descritores locais depende da quantidade de pontos-de-interesse detectados na imagem, o que pode resultar em um número enorme e que varia de imagem para imagem, prejudicando o desempenho dos sistemas de recuperação em grandes bases de dados.

Nos últimos anos a abordagem *Bag-of-Visual-Words* (BoVW), também chamada de *Bag-of-Features* ou *Bag-of-Keypoints*, tem recebido destaque na área de visão computacional como uma modelagem capaz de lidar com o grande e variável número de descritores locais que podem ser extraídos de uma imagem. A abordagem BoVW é uma representação global para analisar os descritores locais extraídos de uma imagem. A ideia é quantizar esses descritores locais em “palavras-visuais” e, então, lidar com cada imagem como uma *coleção-de-palavras*. Desse modo, podem-se utilizar, por analogia, os conceitos e resultados dos Sistemas de Recuperação Textual, podendo aplicar técnicas de indexação escaláveis para possibilitar uma busca rápida no espaço vetorial. A abordagem BoVW é semelhante à abordagem *Bag-of-Words* (BoW) utilizada para representar documentos textuais, entretanto, embora simples e diretamente emprestado da comunidade de recuperação de textos, a abordagem BoVW tem mostrado excelente desempenho não apenas para a tarefa CBIR mas também para outras tarefas de visão, como reconhecimento de objetos, classificação de imagens e anotações.

1.3 Definição do Problema

A abordagem BoVW pode ser dividida em dois passos críticos (Figura 1.1): codificação (*coding*) e sumarização (*pooling*). O passo de codificação discretiza os descritores locais extraídos da imagem em palavras-visuais de acordo com um dicionário visual, e a etapa de sumarização quantifica (sintetiza) as palavras-visuais extraídas em um único vetor-de-características que representará a imagem nos processos de classificação e/ou recuperação. Na abordagem BoVW clássica, chamada de *hard assignment coding*, o passo de codificação simplesmente atribui à cada descritor local extraído da imagem a palavra-visual mais próxima do dicionário visual. Já a fase de sumarização clássica representa as imagens em um histograma de palavras-visuais, que indica a frequência com que cada palavra-visual aparece na imagem.

Figura 1.1: Abordagem clássica da metodologia *Bag-of-Visual-Words*.

Trabalhos frequentemente encontrados na literatura defendem que a escolha das técnicas de codificação e de sumarização tem um enorme impacto sobre o desempenho na representação BoVW (Jiang et al., 2007). Por esse motivo, diversas pesquisas têm sido realizadas com o objetivo de desenvolver métodos de extração de características eficazes e eficientes, de tal maneira que essas características possam quantificar o melhor possível a similaridade perceptual de imagens utilizando a abordagem BoVW. Porém, a utilização da abordagem BoVW ainda continua sendo uma estratégia de caracterização com pontos de desenvolvimento e possibilidades de ajustes, as quais são oportunidades para o avanço de pesquisas. Por exemplo, considere uma situação em que duas imagens A e B apresentam uma grande diferença em apenas uma palavra-visual, e as imagens A e C apresentam diferenças menores, porém em mais de uma palavra-visual, qual imagem B ou C será considerada mais similar à imagem A? Qual é o efeito das variações conjuntas das palavras-visuais no processo da análise da similaridade?

A motivação considerada neste trabalho é que a abordagem BoVW pode ser enriquecida agregando informações úteis e relevantes sobre as palavras-visuais na etapa de sumarização, com o objetivo de complementar a representação semântica das imagens. Várias pesquisas na área se apoiam no fato de que a *codificação da interação e disposição entre as palavras-visuais são informações extremamente importantes que devem ser consideradas na representação das imagens* (Lazebnik et al., 2006; Penatti et al., 2014). Esse fato se baseia na observação de que duas imagens podem ter a mesma quantidade de palavras-visuais, mas não a mesma disposição espacial entre elas. Portanto, o desafio é: como codificar o conjunto das palavras-visuais presentes em uma imagem de uma forma eficaz e eficiente?

Além disso, recuperar imagens usando apenas suas características intrínsecas não é uma abordagem totalmente capaz de extrair e representar o conteúdo semântico presente em uma imagem. Logo, é necessário desenvolver novas metodologias que integrem o contexto e a percepção do usuário no processo de recuperação de imagens.

1.4 Objetivos e Contribuições

Em suma, o principal problema investigado nesta de tese de doutorado pode ser formulado pela seguinte pergunta:

É possível estender a abordagem BoVW para ampliar sua representação semântica em sistemas CBIR?

O objetivo geral deste trabalho foi demonstrar que a abordagem baseada em dicionários visuais e palavras-visuais possui potencial para incorporar descrições e ferramentas que contemplem a semântica da informação presente nas imagens. Desse modo, ao se utilizar BoVW enriquecidos semanticamente, sistemas CBIR podem se aproximar melhor ao esperado pelo usuário. Para que esse objetivo geral fosse alcançado, metas mais específicas foram estabelecidas:

- Desenvolvimento de descritores que agreguem a diversidade de informação presente em uma imagem baseada em dicionários-visuais, tais como a distribuição das palavras-visuais no espaço da imagem;
- Proposta de um arcabouço teórico que permita consultas por similaridade de teor contextual em sistemas CBIR.

As principais contribuições deste trabalho podem ser destacadas nas três seguintes propostas:

- A primeira proposta estende a metodologia BoVW permitindo incluir a informação espacial das palavras-visuais em qualquer tipo de imagem e, deste modo, enriquecendo a tradicional representação baseada em histograma. Serão apresentados três novos formalismos que modelam o conjunto das palavras-visuais detectadas em uma imagem gerando informações complementares na representação final da imagem, por exemplo: a co-ocorrência e disposição global das palavras. Um desses formalismos é a ideia do uso de *frases-visuais* baseadas em *Bunch-of-2-Grams*, que é uma abordagem nova, semanticamente mais robusta e elegante de se analisar a interação das palavras-visuais no espaço da imagem;
- A segunda proposta se refere especificamente à descrição de formas (silhuetas) de objetos, que é uma tarefa computacionalmente desafiadora na área de processamento de imagens. Em reconhecimento de padrões e áreas afins, a silhueta é uma característica amplamente utilizada e explorada em sistemas CBIR. A principal contribuição desse trabalho é a proposta de um novo descritor, chamado de *Bag-of-Saliency-Points* (BoSP), que analisa a similaridade de formas via identificação e caracterização de seus pontos de saliências, inspirado na metodologia

BoVW é adaptado especificamente à descrição de formas. Pontos de saliência são características visuais que tem uma forte inspiração cognitiva e tem uma capacidade de representação compacta, invariante a rotação e translação;

- A terceira contribuição apresenta um novo modelo de consulta por similaridade baseada na ideia de Imagens Dominadas. A motivação é reduzir uma grande coleção de imagens em um conjunto menor que represente a diversidade de imagens da base de dados. Essa é uma estratégia valiosa, pois permite analisar o contexto da aplicação de maneira mais eficiente. A nova medida de similaridade, chamada de *Similarity Based on Dominant Images* ($SimD_{Im}$), faz uso de um conjunto de Imagens Dominantes e atualiza a posição das imagens no *ranking* inicial gerado pelo sistema CBIR, de tal modo que imagens mais similares à imagem-consulta possam aparecer nas primeiras posições desse *ranking* e imagens diferentes no final, que é a expectativa almejada pelo usuário do sistema.

De fato, as diversas contribuições propostas neste trabalho de doutorado permitiram o avanço no estado-da-arte, desenvolvendo metodologias efetivas para caracterizar e quantificar a similaridade entre imagens. As propostas visaram a aproximar a descrição de imagens à subjetividade humana, aprimorando a eficiência de sistemas CBIR, retornando imagens mais próximas ao contexto esperado pelo usuário, e, desse modo, permitindo reduzir de maneira eficaz e eficiente a lacuna semântica existente entre a percepção humana e a descrição computacional.

1.5 Organização do documento

O texto desta tese está organizado da seguinte maneira:

Capítulo 2 : é apresentado o referencial teórico sobre a tarefa da Recuperação de Imagens Baseada em Conteúdo, mostrando a arquitetura clássica de um sistema CBIR (*Content Based Image Retrieval*), como pode ser realizada a análise da similaridade entre imagens, métricas e tipos de descritores de imagens;

Capítulo 3 : são descritas as etapas para se descrever uma imagem em palavras-visuais e como representar o conjunto dessas palavras em *Bag-of-Visual-Words*, listando as técnicas, desafios e caminhos para o avanço de pesquisas nessa área, juntamente com algumas avaliações experimentais realizadas analisando a combinação de diferentes técnicas e o impacto delas na recuperação de imagem;

Capítulo 4 : são propostas três novas técnicas que permitem enriquecer a representação semântica de uma imagem descrita em palavras-visuais, estendendo a abordagem clássica baseada em histograma da modelagem *Bag-of-Visual-Words*;

Capítulo 5 : é apresentado um novo descritor de formas baseado em pontos-de-saliências com uma modelagem inspirada na abordagem de dicionários visuais, porém usando estratégias específicas e adaptadas para a descrição de formas;

Capítulo 6 : é apresentada uma nova metodologia de consulta por similaridade que agrega a informação contextual da coleção de imagens na recuperação, a fim de obter um *ranking* final semanticamente mais próximo ao esperado pelo usuário;

Capítulo 7 : são apresentadas as conclusões, artigos publicados e os trabalhos futuros.

Recuperação de Imagens Baseada em Conteúdo

2.1 Considerações Iniciais

Os sistemas computacionais projetados para realizar a Recuperação de Imagens baseado em Conteúdo são denominados CBIR (CBIR - *Content Based Image Retrieval*), e eles são de interesse principal neste trabalho. Tais sistemas tem sido alvo de pesquisas desde a década de 80, uma vez que um grande conjunto de imagens digitais vem sendo gerado e armazenado em bancos de dados. Os primeiros sistemas CBIR começaram a ser desenvolvidos na década de 80, apresentado por (Chang e Fu, 1980). Porém, foi na década de 90, que começaram a surgir sistemas CBIR mais relevantes como o QBIC (*Query By Image Content*), desenvolvido por (Flickner et al., 1995), que era um dos mais avançados e conhecidos sistemas de recuperação baseado em conteúdo de sua época. O QBIC utiliza os principais aspectos da imagem: cor, forma e textura para efetuar a recuperação de imagens. Histograma de cores é utilizado na recuperação baseada em cor e descritores simples e momentos são empregados na recuperação baseada em forma e textura. Além disso, uma estrutura de indexação ou método de acesso pode ser empregada para trazer rapidez à busca, principalmente se forem consideradas grandes bases de imagens.

Um sistema CBIR ideal é aquele capaz de encontrar de maneira eficiente imagens perceptivamente similares com base em uma coleção de dados, portanto, a detecção e descrição das características de uma imagem são tarefas básicas para a recuperação de imagens. A extração de características robustas e eficientes podem melhorar significativamente a eficácia da tarefa de recuperação de imagens em sistemas CBIR. Logo, os resultados obtidos por esses sistemas dependem fortemente da escolha das características

utilizadas para representar uma imagem. Desse modo, o ponto-chave desses sistemas é a obtenção de dados discriminantes que corresponderiam a um possível julgamento humano da similaridade entre a imagem de consulta e as outras pertencentes ao banco de imagens. Uma linha de pesquisa bastante explorada é tentar minimizar o *gap*-semântico existente entre o poder de interpretação limitado das máquinas e a subjetividade humana. Porém, definir as características de uma imagem pode se tornar uma tarefa árdua, uma vez que essas características dependem do domínio de aplicação.

Neste capítulo será apresentada uma visão geral do funcionamento de um sistema de recuperação de imagens baseado em conteúdo, bem como as principais métricas utilizadas para analisar a similaridade entre as características extraídas das imagens e as medidas para avaliar o desempenho de um sistema CBIR.

2.2 Arquitetura de um Sistema CBIR

A busca de imagens por meio de sistemas CBIR é usualmente executada utilizando uma imagem de consulta com a finalidade de recuperar as imagens mais semelhantes à imagem referência (ou de consulta). Esse método é chamado de Consulta-por-Exemplo (*Query by Example*) ou QBE. A recuperação começa quando o sistema apresenta uma seleção inicial de imagens de referência através de uma interface. O sistema pode apresentar alguma imagem representativa ou o usuário pode escolher alguma aleatoriamente. Essa imagem selecionada é usada para recuperar as imagens mais semelhantes a ela dentro de um banco de dados.

O funcionamento de um sistema CBIR simples pode ser ilustrado pelo *framework* presente na Figura 2.1: dada uma imagem de consulta, aplica-se um algoritmo de extração de características sobre essa imagem e sobre todas as imagens da base de dados; então é utilizada uma “função de distância” para medir a similaridade entre as características extraídas da imagem de consulta e as extraídas das imagens da base de dados. Por fim o sistema retorna as imagens mais próximas (mais similares) à imagem de consulta e envia para um módulo de interface onde serão visualizadas pelo usuário.

Geralmente, o processo de extração dos vetores-de-características é realizado uma única vez para cada imagem da base de dados, sendo realizado de maneira *offline*. Os vetores armazenados são usados posteriormente no processamento de consultas.

2.2.1 Estruturas de Indexação

O uso adequado de estruturas de indexação é um aspecto importante na implementação de um sistema CBIR. Este aspecto está relacionado à utilização de estruturas apropriadas que possam agilizar a realização de consultas por similaridade, minimizando o número de comparações/cálculos necessários para executar uma consulta. As estruturas de

Figura 2.1: Metodologia de um sistema de recuperação de imagem por conteúdo.

indexação/métodos de acesso são ferramentas fundamentais que permitem aos sistemas de recuperação de imagens armazenar e recuperar eficientemente um grande volume de informação.

Os Métodos de Acesso Métricos (*Metric Access Methods* - MAM) são os mais adequados em sistemas de consulta por similaridade em espaços métricos que englobam tanto dados espaciais com dimensão definida quanto dados adimensionais. Os MAM são estruturas de indexação baseadas em distância que utilizam exclusivamente funções de distância para organizar os objetos na base de dados. A ideia básica consiste na escolha de um objeto arbitrário central e na aplicação de uma função de distância para dividir os demais objetos em vários subconjuntos. Dessa maneira, a estrutura de indexação é construída executando-se esse mesmo procedimento, recursivamente, para cada subconjunto não vazio.

Na maioria dos sistemas de recuperação, as imagens no banco de dados são indexadas, de acordo com seus vetores-de-características, pelo uso de estruturas como *M-tree* (Ciaccia et al., 1997), *Slim-tree* (Traina et al., 2002), ou outros MAMs.

2.2.2 Gap Semântico e a Realimentação de Relevância

Ao tentar recuperar uma imagem em um sistema de recuperação de imagens, o usuário depara-se com o problema da quantidade excessiva de dados irrelevantes, seja por dificuldade da escolha de uma imagem adequada para consulta ou pela falta de descritores apropriados que consiga incorporar a subjetividade da percepção visual. Tal problema está relacionado com o *gap* semântico, que é a diferença entre características de baixo nível extraídas da imagem, e o alto nível da informação que o usuário necessita. O objetivo da redução do *gap* semântico é estabelecer uma ligação entre a expectativa do usuário e as características extraídas, usualmente de modo automático, da imagem.

A técnica de realimentação de relevância (*relevance feedback*), tem sido utilizada com o intuito de incorporar a subjetividade da percepção visual de usuários à recuperação de imagens por conteúdo. Essa técnica possibilita ao usuário informar quais as imagens ele

considera relevantes dentro de um conjunto de imagens retornado pelo sistema. A ideia da técnica de realimentação de relevância é “aprender” a vontade do usuário durante um determinado número de iterações, nas quais o usuário indica os resultados que são considerados por relevantes (positivos) e os irrelevantes (negativos). Dessa maneira, o sistema poderá retornar imagens cada vez mais similares à vontade do usuário, aprendendo o conceito estabelecido por ele.

Um aspecto importante para a definição de um mecanismo de realimentação de relevância é o algoritmo de aprendizado. Em alguns trabalhos, o aprendizado consiste em estimar o vetor-de-característica que melhor representa o padrão de consulta. Outros atribuem pesos para cada posição do vetor-de-características e para cada descritor utilizado. Assim, o aprendizado consiste em estimar esses pesos que melhor representem a percepção visual do usuário.

2.3 Análise da Similaridade

Os sistemas de recuperação tradicionais realizam suas consultas por igualdade e ordem. Porém, as aplicações que lidam com dados complexos requerem a realização de consultas que realizem busca por objetos da base que sejam similares a um objeto de consulta, de acordo com uma certa medida de similaridade e baseada em uma função de distância.

Para conseguir uma recuperação de imagens eficaz, que satisfaça às expectativas do usuário, as características extraídas da imagem devem ser combinadas com uma função de distância. Essa função é definida de forma a calcular o grau de similaridade entre dois objetos, retornando um valor que quantifica o quão dissimilar eles são. Esse valor é maior que zero e é inversamente proporcional ao nível de similaridade das imagens, ou seja, para objetos menos similares esse valor é grande e para objetos muito similares ele é próximo de zero.

2.3.1 Propriedades Métricas

As funções de distância, empregadas pelas técnicas de recuperação por conteúdo, são definidas dentro de um espaço métrico e podem ser formalmente descritas da seguinte maneira: dado um conjunto $\mathcal{O}$, uma função de distância métrica em $\mathcal{O}$ é uma função $d : \mathcal{O} \times \mathcal{O} \rightarrow \mathbb{R}$ que possui as seguintes propriedades:

1. Simetria: $d(o_q, o_r) = d(o_r, o_q)$
2. Não negatividade: $d(o_q, o_r) > 0$ se $o_q \neq o_r$ e $d(o_q, o_r) = 0$ se $o_q = o_r$
3. Desigualdade triangular: $d(o_q, o_p) \leq d(o_q, o_r) + d(o_r, o_p)$

em que o_p, o_q e $o_r \in \mathcal{O}$.

É desejável que uma medida de similaridade possa satisfazer todas as três propriedades. Se essas propriedades são requeridas ou não, irá depender da aplicação.

2.3.2 Funções de Distância

Existem muitas funções de distância que podem ser aplicadas para medir a similaridade entre dois vetores-de-características. Algumas das funções de distância mais comuns, usadas por descritores de imagens, são:

- **Distância de Minkowski:** Considerando dois pontos O e O' , pertencentes a um espaço vetorial n -dimensional, ou seja, $o = \{o_1, \dots, o_n\}$ e $o' = \{o'_1, \dots, o'_n\}$, a distância de Minkowski, L_p , é dada por:

$$L_p(O, O') = \sqrt[p]{\sum_{i=1}^n |o_i - o'_i|^p} \quad (2.1)$$

em que p é um parâmetro que modifica a distância de Minkowski resultando em outras distâncias conhecidas. Se $p = 2$ tem-se a distância Euclidiana, se $p = 1$ tem-se a distância City-block. Caso $p \geq 1$ tem-se uma distância métrica, e caso $p < 1$ a desigualdade triangular não é obedecida o que resulta na inexistência de uma distância métrica.

Uma variante da família de distâncias Minkowski é a distância Minkowski ponderada L'_p , a qual atribui pesos distintos no cálculo da distância, levando-se em consideração que determinados atributos podem exercer maior ou menor relevância em relação ao contexto que está sendo analisado. Essa distância ponderada é formalmente definida como:

$$L'_p(O, O') = \sqrt[p]{\sum_{i=1}^n w_i |o_i - o'_i|^p} \quad (2.2)$$

em que w_i é um elemento do vetor de ponderação $w = (w_1, \dots, w_n)$. Portanto, para diferentes contextos podem ser estipulados diferentes vetores de ponderação.

- **Distância χ^2 :** esta função valoriza as grandes discrepâncias entre os dois vetores, tal que atributos muito semelhantes são considerados mais próximos do que quando há diferença muito grande ou muito baixa entre eles. Essa distância é definida como:

$$\chi^2(O, O') = \sum_{i=1}^n \frac{(o_i - m_i)^2}{m_i} \quad (2.3)$$

em que, $m_i = (o_i + o'_i)/2$.

- **Intersecção de Histogramas:** é normalmente utilizada para medir a distância entre histogramas. Esta função é interessante devido ao fato de ser possível tratar intervalos parciais, quando a área de um histograma é menor do que a do outro. Por outro lado, quando as áreas dos dois histogramas são iguais, a Intersecção de Histogramas transforma-se na função L_1 . Ela é definida por:

$$IH(\mathcal{O}, \mathcal{O}') = 1 - \frac{\sum_{i=1}^n \min(o_i, o'_i)}{\sum_{i=1}^n o_i} \quad (2.4)$$

- **Canberra:** é uma versão ponderada da função L_1 , definida por:

$$d_{Canberra}(\mathcal{O}, \mathcal{O}') = \sum_{i=1}^n \frac{|o_i - o'_i|}{|o_i| + |o'_i|} \quad (2.5)$$

- **Distância Cosseno:** é uma medida de distância entre dois vetores de um espaço interno do produto que mede o cosseno do ângulo entre eles, definida por:

$$d_{Cosseno}(\mathcal{O}, \mathcal{O}') = \frac{\sum_{i=1}^n o_i o'_i}{\sqrt{\sum_{i=1}^n (o_i)^2} \sqrt{\sum_{i=1}^n (o'_i)^2}} \quad (2.6)$$

- **Distância de Hausdorff:** Em muitas aplicações, nem todos os pontos de um conjunto $\mathcal{O}$ precisam ter um ponto correspondente em um outro conjunto $\mathcal{O}'$. Tipicamente, esses dois conjuntos $\mathcal{O}$ e $\mathcal{O}'$ são de tamanho diferentes, sendo assim não existe uma correspondência 1-1 entre todos os pontos. Nesse caso, a medida de similaridade frequentemente usada é a distância de Hausdorff.

Dados dois conjuntos de pontos $\mathcal{O}$ e $\mathcal{O}'$, a distância de Hausdorff entre esses conjuntos é dado por:

$$H(\mathcal{O}, \mathcal{O}') = \max\{\max_{o \in \mathcal{O}} \min_{o' \in \mathcal{O}'} d(o, o'), \max_{o' \in \mathcal{O}'} \min_{o \in \mathcal{O}} d(o, o')\} \quad (2.7)$$

em que $d(o, o')$ é a distância entre os elementos de $\mathcal{O}$ e $\mathcal{O}'$ e normalmente se utiliza distância Euclidiana.

Embora a distância de Hausdorff permita o casamento parcial entre dois conjuntos, ela não possui invariância à mudança de escala, rotação ou translação, fato que restringe bastante o seu uso.

2.3.3 Tipos de Consultas por Similaridade

Existem basicamente dois tipos de consultas por similaridade: as consultas por abrangência (*Range query* - R_q), cujo foco está centrado em todos os objetos da base de dados que sejam similares ao objeto de consulta até no máximo um certo limiar; e

as consultas aos k -Vizinhos mais Próximos (k -Nearest Neighbor query - kNNq), que está relacionado à recuperação dos k objetos mais similares ao objeto de consulta.

Seja um objeto de consulta $o_q \in \mathbb{O}$, um conjunto de objetos $\mathcal{O} \subseteq \mathbb{O}$ e uma função de distância $d()$, esses dois tipos de consultas podem ser definidos formalmente como:

- **Consulta por Abrangência:** dado uma distância máxima de busca ξ , a consulta por abrangência $R_q(o_q, \xi)$ recupera todos os objetos de $\mathcal{O}$ que estejam a uma distância de o_q de no máximo ξ . O subconjunto resposta resultante $\mathcal{O}' \subseteq \mathcal{O}$ é $\{o_i \in \mathcal{O} \mid d(o_q, o_i) \leq \xi\}$.
- **Consulta aos k -Vizinhos mais Próximos:** dado um número inteiro k , a consulta k-NNq(o_q, k) recupera os k objetos mais próximos de o_q em $\mathcal{O}$. O subconjunto resposta resultante $\mathcal{O}' \subseteq \mathcal{O}$ é $\{o_q \in \mathcal{O} \mid |\mathcal{O}'| = k \text{ e } \forall o_r \in \mathcal{O}', \forall o_j \in [\mathcal{O} - \mathcal{O}'], d(o_q, o_r) \leq d(o_q, o_j)\}$.

2.4 Avaliação do Desempenho de um Sistema CBIR

Quando um usuário realiza uma consulta através de um sistema CBIR, ele espera que este sistema recupere apenas imagens que sejam do seu interesse (imagens relevantes). Portanto, é possível verificar se um sistema CBIR satisfaz à expectativa do usuário se o mesmo está recuperando uma grande quantidade de imagens relevantes e também se possui precisão na recuperação, ou seja, se não fornece muitas imagens irrelevantes para recuperar uma quantidade satisfatória de documentos relevantes.

O que se deseja na realidade é estimar se o sistema CBIR está obtendo imagens pertinentes à consulta. Sendo assim, é preciso saber de antemão quais são as imagens consideradas relevantes pelo usuário.

Considerando uma base de imagens à qual se submete uma consulta, vários conjuntos de imagens estão envolvidos. Dentre eles os mais importantes são: o conjunto de imagens pertinentes à consulta (conjunto de elementos relevantes R) e o conjunto de imagens recuperadas (conjunto resposta A). Dentro desse último tem-se imagens relevantes e não relevantes, sendo que o conjunto de maior interesse é o de imagens relevantes recuperadas Ra . A Figura 2.2 ilustra a relação entre estes conjuntos.

2.4.1 Curva de Precisão vs. Revocação

Dentre as diversas medidas para avaliação do desempenho de um sistema CBIR, a medida de precisão e revocação são as principais. Dada uma consulta, a *precisão* é definida como a fração entre o número de imagens relevantes recuperadas, $|Ra|$, e o número de imagens recuperadas, $|A|$:

Figura 2.2: Conjuntos de imagens avaliadas na Recuperação de Imagens baseada em Conteúdo.

$$Pr(Ra) = \frac{|Ra|}{|A|} \quad (2.8)$$

Por outro lado, a *revocação* é definida como a fração entre o número de imagens relevantes recuperadas, $|Ra|$, e o número total de imagens relevantes na coleção, $|R|$:

$$Re(Ra) = \frac{|Ra|}{|R|} \quad (2.9)$$

Essas duas medidas estão intimamente ligadas, pois dependem da quantidade $|Ra|$ de imagens relevantes recuperadas. Em geral, aumentando-se a revocação a precisão diminui. Isso ocorre porque à medida que imagens são inseridas na resposta à uma dada consulta, aumenta-se a quantidade de imagens relevantes recuperadas, contudo juntamente com essas imagens relevantes, imagens não relevantes são inseridas em maior proporção que as relevantes, produzindo o efeito da redução da precisão. Uma maneira de facilitar a análise entre as mesmas é traçar o gráfico (curva) da Precisão em relação à Revocação. Quanto maior a área da curva, melhor o resultado.

2.4.2 Precisão Média

A maneira mais comum de resumir a curva de Precisão *vs.* Revocação em um único valor é utilizando a precisão média (Deselaers et al., 2008). A precisão média $\overline{Pr}(q)$ para uma única consulta q é a média da precisão após cada imagem relevante recuperada:

$$\overline{Pr}(q) = \frac{1}{|R|} \sum_{i=1}^{|R|} Pr(i) \quad (2.10)$$

em que $|R|$ é o total de imagens relevantes para a consulta q .

Usualmente os sistemas de recuperação de informação são avaliados para várias consultas distintas. Para avaliar o desempenho da recuperação sobre todas as consultas de teste, a taxa média de precisão (*mAP* - *mean Average Precision*) é adotada:

$$mAP = \frac{1}{|C|} \sum_{q \in C} \overline{Pr}(q) \quad (2.11)$$

em que $|C|$ é a quantidade de imagens da base de dados.

2.5 Descritores de Imagens

Imagens devem ser representadas por um *extrator de características* com a finalidade de gerar um vetor-de-características que as represente numericamente e cujos valores serão usados nos processos de indexação, recuperação e/ou classificação.

Um vetor-de-características $\vec{v}$ é um ponto no espaço $\mathbb{R}^n$: $\vec{v} = (\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n)$, em que n é a dimensão do vetor. Basicamente um descritor tem o objetivo de caracterizar as informações presentes nas imagens, de tal maneira que essas informações sejam compatíveis com a intuição humana. Esse vetor sintetiza as propriedades da imagem de acordo com os elementos escolhidos e pode ter qualquer dimensão. Diferentes tipos de vetores-de-características podem necessitar de funções de similaridade distintas.

2.5.1 Descritores Globais

A maioria dos sistemas CBIR recuperam imagens a partir das características de baixo nível da imagem, tais como cor, textura e forma. Quase todos estes sistemas se baseiam na premissa de que as imagens podem ser caracterizadas por *descritores globais* para recuperar imagens semelhantes em um banco de dados. Um descritor global é constituído por características computadas considerando a imagem toda.

Geralmente, as características de cor são facilmente obtidas diretamente a partir da intensidade dos pixels, por exemplo, histograma de cores. Por outro lado, pode-se utilizar matrizes de co-ocorrências para caracterizar texturas. A noção de textura geralmente se refere à presença de um padrão espacial, que tem algumas propriedades de homogeneidade. Em particular, a homogeneidade não pode resultar da presença de apenas uma cor única para as regiões, mas exige interação de várias cores. Descritores de texturas são mais utilizados para caracterizar imagens médicas, pois esses descritores permitem descrever detalhes finos dentro da estrutura da imagem.

Descritores de formas permitem caracterizar objetos presentes na imagem, portanto exige-se que o objeto já esteja segmentado da imagem. O problema de segmentação pode ser visto como o principal obstáculo para o uso dos descritores de formas. No contexto médico, a informação da forma é um dos fatores mais fortes na detecção de uma

determinada doença/lesão ou na compreensão de sua evolução. No entanto, a maioria dos atuais sistemas CBIR médicos não exploram todo o potencial da informação de forma. A maioria utiliza descrições simples que são incapazes de captar a semântica necessária para medir a granularidade de um objeto.

A Tabela 2.1 lista alguns dos descritores utilizados na literatura. A matriz de co-ocorrência (Haralick et al., 1973) é uma das técnicas mais tradicionais para codificar o relacionamento espacial entre níveis de cinza de uma imagem. Com ela pode-se extrair os descritores de Haralick, que é um descritor de textura bastante utilizado na área médica. Outro exemplo de descritor de textura são os coeficientes de Gabor (Manjunath e Ma, 1996). Já para analisar cores, o histograma de cores continua sendo o descritor mais utilizado. E para descrever formas, alguns trabalhos como o de (Rahman et al., 2011), utilizam um histograma dos ângulos das regiões de bordas da imagem. As bordas são extraídas usando o método de Canny e os ângulos de cada pixel de borda são quantizados em 72 *bins* com valores incrementais de 5 graus cada.

Tabela 2.1: Alguns dos descritores mais utilizados em sistemas CBIR, em suas configurações tradicionais.

Categoria	Descritor	Tamanho do vetor-de-característica
cor	Histograma	255
cor	Histograma em blocos	qtde de blocos * 255
cor	Momentos	7
textura	Haralick	14
textura	Gabor	48
textura	Wavelets	128
forma	Fourier	128
forma	Histograma de ângulos	72

2.5.2 Descritores Locais

Imagens podem representar uma cena que consiste em uma coleção de objetos diferentes (ou regiões), e, portanto, uma descrição individualizada dessas regiões permitem uma melhor representação do conteúdo da imagem. A solução consiste em separar essas diferentes regiões da imagem utilizando algum algoritmo de segmentação e, em seguida, utilizar as características apropriadas calculadas em cada região da imagem para descrevê-las. Estas abordagem de analisar uma imagem constituem os chamados *descritores locais*.

Descritores locais, portanto, necessitam de duas fases distintas: detecção e descrição. A primeira fase consiste em encontrar um conjunto de pontos-de-interesse ou regiões salientes que sejam invariantes às transformações geométricas. A segunda fase utiliza algum ferramental para caracterizar de maneira robusta as regiões detectadas na primeira fase.

A vantagem da descrição global sobre a descrição local é a facilidade na análise da similaridade entre imagens, pois ela gera apenas um vetor-de-características representativo para toda a imagem. A descrição local, entretanto, gera para cada região-de-interesse um vetor-de-características, logo para analisar a similaridade entre duas imagens descritas localmente é necessário alguma abordagem que consiga medir a distância entre vetores com tamanho diferentes.

Detecção das Regiões-de-Interesse

A detecção das regiões-de-interesse é a extração de padrões locais representativos da imagem. Esses padrões são os elementos básicos que compõem a imagem, de tal maneira que o conjunto desses padrões é utilizado para medir a similaridade entre duas imagens.

Existem diversas técnicas na literatura desenvolvidas para a detecção automática de regiões-de-interesse em uma imagem. Tal como ilustrado na Figura 2.3, essas técnicas se dividem em duas abordagens:

- **aleatória:** existem trabalhos que utilizam uma *abordagem aleatória* para a extração de regiões-de-interesse na imagem, por exemplo, usando uma grade-regular (Marée et al., 2005). A ideia dessa abordagem é dividir a imagem em regiões homogêneas, , por exemplo como uma grade reticulada, sem qualquer tipo de semântica envolvida nessa divisão;
- **pontos-de-interesse:** os *pontos-de-interesse* de uma imagem são pontos caracterizados por terem alguma semântica envolvida. Por exemplo, a figura mostra os pontos de quina (*corner points*) da imagem.

O desafio dos trabalhos que utilizam a abordagem baseada em *pontos-de-interesse* está na localização de pontos que sejam invariantes a transformações geométricas, iluminação e ruído, possibilitando a localização dos mesmos pontos em imagens distintas. Para isso, na literatura existem diversas técnicas propostas, cada uma enfatizando diferentes aspectos de invariância na imagem, resultando em métodos com propriedades e tamanhos de amostras distintas.

Em geral, os mais populares detectores de pontos-de-interesse utilizados na literatura são: Laplaciano da função Gaussiana (LoG) (Lindeberg, 1998), Harris Laplace e Affine (Mikolajczyk e Schmid, 2004), Hessian Laplace e Affine (Mikolajczyk et al., 2005) e a Diferença de Gaussianas (DoG) (Lowe, 1999). A maioria desses detectores é baseada em derivadas: por exemplo, o detector Hessian proposto por (Beaudet, 1978), procura por regiões na imagem que apresentam mudanças bruscas no valor do gradiente, calculado pela seguinte matriz de autocorrelação:

$$Hessian(I) = \begin{bmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{bmatrix} \quad (2.12)$$

Figura 2.3: Metodologia para detectar as regiões-de-interesse de uma imagem: grade-regular ou pontos-de-interesse.

em que I_{xx} , I_{xy} e I_{yy} correspondem às segundas derivadas da imagem $I(x, y)$, com relação às coordenadas x e y . A proposta do detector Hessian é selecionar pontos cujo $\det(\text{Hessian}(I))$ seja maior que um determinado valor θ , em que:

$$\det(\text{Hessian}(I)) = I_{xx} \cdot I_{yy} - I_{xy}^2 \quad (2.13)$$

O detector de Hessian, além de ser bastante simples, gera bons resultados em imagens simples e sem muito ruído. Porém, para imagens muito ruidosas e com textura densa o detector seleciona muitos pontos irrelevantes e que prejudicam na qualidade da representação da imagem. Para resolver esse problema, a proposta do detector de Harris é utilizar uma função Gaussiana convoluída com uma matriz de autocorrelação:

$$\mu(x, y, \sigma) = \sigma^2 \cdot g(x, y, \sigma) * \begin{bmatrix} I_x^2(x, y) & I_x I_y(x, y) \\ I_x I_y(x, y) & I_y^2(x, y) \end{bmatrix} \quad (2.14)$$

em que I_x e I_y são as respectivas derivadas nas direções x e y da imagem $I(x, y)$; e $g(x, y, \sigma)$ é uma função Gaussiana circular com desvio padrão σ :

$$g(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{\left(\frac{-x^2-y^2}{2\sigma^2}\right)} \quad (2.15)$$

Os pontos-de-interesse são dados por:

$$\text{Harris} = \det(\mu) - \alpha \cdot \text{trace}^2(\mu) \quad (2.16)$$

em que α é uma constante.

A função Gaussiana é amplamente utilizada na área de processamento de imagens como um filtro passa-baixa, cuja aplicação tende a borrar a imagem e eliminar ruídos. Ela também é utilizada pelo detector de Diferenças de Gaussianas (DoG), que é o detector de pontos-de-interesse utilizado pelo descritor SIFT (*Scale Invariant Feature Transform*) (Lowe, 1999). A proposta do detector DoG é a localização de pontos locais realizando a convolução da imagem com uma função Gaussiana contínua em diferentes escalas. O espaço da escala é definido como uma função $L(x, y, \sigma)$, obtido pelo resultado da convolução da imagem de entrada $I(x, y)$ com a função Gaussiana de escala variável $g(x, y, \sigma)$:

$$L(x, y, \sigma) = g(x, y, \sigma) * I(x, y) \quad (2.17)$$

em que $*$ é o operador de convolução. Os pontos locais candidatos são computados pela diferença de duas escalas próximas separadas por um fator constante multiplicativo k :

$$\text{DoG}(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (2.18)$$

ou seja, $\text{DoG}(x, y, \sigma)$ representa a diferença entre imagens submetidas a um filtro Gaussiano em escala σ e $k\sigma$. Um ponto-de-interesse é um ponto de máximo ou mínimo local de $\text{DoG}(x, y, \sigma)$, quando comparado com seus 8-vizinhos na imagem corrente e 9-vizinhos na escala acima e abaixo. Além disso, pontos que possuem baixo contraste ou não estão próximos à uma quina são rejeitados.

Além das técnicas baseadas em derivadas, há também o detector SUSAN (*Smallest Univalued Segment Assimilating Nucleus*), proposta por (Smith e Brady, 1995), que é um exemplo de método eficiente com base em operadores morfológicos. A ideia por trás deste detector é usar uma máscara circular em todos os pontos da imagem. A região dessa máscara é dada por M , cada ponto da imagem dentro dessa máscara é dado por $m \in M$ e seu núcleo (centro) é representado pelo ponto m_0 . Essa máscara M é percorrida por toda imagem e cada ponto $m \in M$ é comparado com o núcleo m_0 (centro) usando a seguinte função:

$$c(m) = e^{-\left(\frac{I(m) - I(m_0)}{t}\right)} \quad (2.19)$$

em que t é o raio da máscara circular e $I(m)$ o valor do nível de cinza do ponto m . A área do SUSAN é calculada por:

$$A(M) = \sum_{m \in M} c(m) \quad (2.20)$$

se $A > g$, em que g é um “*threshold* geométrico”, então o ponto $m = g - A(M)$. Por fim, é realizado uma supressão não-máxima para encontrar e eliminar pontos não-relevantes.

A quantidade de regiões-de-interesse em uma imagem pode variar em relação ao tipo de detector escolhido e também em relação à imagem a ser descrita, como apresentado no trabalho de (Mikolajczyk e Schmid, 2005) e ilustrado pela Figura 2.4. Devido a esta quantidade de informações geradas, a utilização de alguma técnica que reduza o número de dimensões pode tornar-se vantajosa. Neste caso o PCA-SIFT (*Principal Components Analysis-SIFT*) (Ke e Sukthankar, 2004) reduz a dimensão do vetor-de-características, diminuindo a quantidade de informação a ser indexada. O PCA é uma técnica estatística capaz de identificar padrões em dados e expressá-los de tal forma que suas diferenças e similaridades sejam realçadas, além de possibilitar a compressão dos dados pela redução do número de dimensões, sem grande perda de informação.

Figura 2.4: Pontos-de-interesse detectados pelas técnicas: (a) Harris , (b) Diferença de Gaussianas (DoG), (c) SUSAN.

Descritor SIFT

O descritor SIFT, além de detectar pontos-de-interesse na imagem, descreve cada ponto utilizando a informação da magnitude e orientação do gradiente da região ao redor de cada ponto-de-interesse. As magnitudes $m(x, y)$ do gradiente e orientações $\theta(x, y)$ são calculadas ao redor do ponto-de-interesse $L(x, y)$ por:

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \quad (2.21)$$

$$\theta(x, y) = \tan^{-1}\left(\frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)}\right) \quad (2.22)$$

Cria-se então um histograma das orientações sobre cada região e, ao somar uma orientação ao histograma, a magnitude é ponderada por uma gaussiana, de forma que orientações referentes a pontos mais distantes tenham menor peso. O descritor é formado por um vetor contendo as magnitudes de todas as orientações dos histogramas, como

pode ser visto na Figura 2.5. No algoritmo são utilizadas matrizes 4x4, resultando em um vetor de 128 elementos (4x4 histogramas de 8 orientações cada um) para cada ponto de interesse.

Figura 2.5: O descritor SIFT é criado computando-se a magnitude do gradiente e a orientação da região em volta do ponto de interesse. Os gradientes são ponderados por uma janela Gaussiana, indicada pelo círculo. As amostras são acumuladas em histogramas de orientação (8 direções) para cada sub-região, que no exemplo mostrado à direita tem tamanho 4x4. O tamanho das setas corresponde à soma das magnitudes dos gradientes próximos àquela direção. Imagem traduzida de (Lowe, 2004).

Entretanto, descritores recentes como o SURF (*Speeded Up Robust Features*), proposto por (Bay et al., 2008), têm mostrado um desempenho similar ao SIFT com tempo de execução mais curto. No SURF são utilizadas técnicas que visam melhorar o tempo gasto na detecção dos pontos-de-interesse e na descrição deles. No processo de localização dos pontos-de-interesse é utilizado o conceito de imagens integrais que é uma representação da imagem que torna convoluções de filtros mais simples e também é utilizada a matriz Hessiana com derivadas de segunda ordem das gaussianas de um ponto, chamada de Hessiana rápida. O processo de descrição dos pontos-de-interesse detectados se assemelha ao método usado pelo SIFT, em que é calculada a intensidade dos pixels vizinhos ao ponto-de-interesse. Uma das diferenças é que a técnica usada pelo SURF para descobrir a orientação de um pixel vizinho é a transformada de Haar, que é mais rápida que o cálculo de gradiente. Outra diferença do SURF é o tamanho do descritor (64 elementos), metade da dimensão do descritor SIFT.

2.6 Considerações Finais

Neste capítulo foi apresentada uma visão geral do funcionamento de um sistema de recuperação de imagem baseado em conteúdo, em inglês CBIR (*Content Based Image Retrieval*). Essa abordagem de recuperar informação é de grande interesse para diversas aplicações, uma vez que um grande conjunto de imagens digitais vem sendo gerado e armazenado em bancos de dados. Foi visto que um CBIR ideal é aquele capaz de encontrar

de maneira eficiente imagens perceptivamente similares com base em uma coleção de dados. Por isso, é necessário que ele faça uso de um *descriptor* que seja capaz de caracterizar as informações presentes nas imagens, de tal maneira que elas sejam compatíveis com a intuição humana.

O primeiro passo para recuperar uma imagem é extrair informações das características intrínsecas contidas nessa imagem, que podem ser cor, forma ou textura, gerando um vetor-de-características. Por fim, a similaridade é feita através de uma função de distância entre esses vetores, retornando um valor real positivo que indica o grau de similaridade entre eles. Foram apresentadas várias funções de distância que podem ser aplicadas para medir a similaridade entre dois vetores, como também as propriedades que essa função deve satisfazer para ser considerada uma função de distância métrica e os tipos de consulta por similaridade.

A descrição de uma imagem pode ser realizada de modo global ou local. A diferença entre essas duas abordagens está na maneira de analisar a imagem: a primeira analisa a imagem como um todo, já a segunda analisa separadamente a imagem em regiões-de-interesse. Computacionalmente, a modelagem global é mais “barata” do que a modelagem local, pois a global gera apenas um único vetor-de-características por imagem, enquanto que a local pode gerar vários vetores, desse modo o custo para analisar a similaridade é mais alto. Para contornar esse problema, no próximo capítulo é apresentada a modelagem *Bag-of-Visual-Words* que representa as características locais extraídas de uma imagem em um único vetor-de-características de tamanho fixo, o que permite o uso de funções de distâncias rápidas, reduzindo o custo computacional.

Descrevendo Imagens em Palavras-Visuais

3.1 Considerações Iniciais

Assim como diz o antigo provérbio “Uma imagem vale mais que mil palavras...”, na área da Computação pode-se afirmar que uma imagem possui uma variedade enorme de dados a serem interpretados tanto quantitativa quanto qualitativamente. O grande desafio é que, para cada domínio específico do conhecimento humano (Medicina, Física, Biologia...), uma imagem pode ser definida e tratada sob diferentes aspectos. Entretanto, computacionalmente existem três restrições únicas que devem ser consideradas na busca de imagens por conteúdo em uma grande base de dados: a precisão do resultado, sua eficiência e a memória usada na representação da imagem. As duas últimas restrições estão interligadas, uma vez que a eficiência da busca é dada pela quantidade de memória necessária para representar a imagem.

Analisar a similaridade de imagens através apenas de suas características locais não é uma tarefa fácil e computacionalmente barata. Descritores locais exigem mais memória e funções de distâncias “especiais” para medir a similaridade na tarefa de recuperação, já que a cardinalidade das características locais podem variar de imagem para imagem. Pesquisas tem sido realizadas para contornar tal problema e permitir sumarizar os descritores locais em apenas um único vetor-de-características e, assim, reduzir o custo computacional da análise da similaridade, tal como a abordagem *Bag-of-Visual-Words* que é baseada na descrição da imagem em *palavras-visuais*.

Neste capítulo será apresentada em detalhes a modelagem *Bag-of-Visual-Words* com definições, discussões e avaliações experimentais próprias, mostrando as técnicas adotadas

na literatura e os desafios computacionais do modelo. Aqui será abordado em detalhes a análise da influência de quatro fatores necessários na implementação da abordagem BoVW:

- o tamanho do dicionário visual;
- o detector dos pontos-de-interesse;
- o tipo de codificação;
- a função de distância usada para medir a similaridade.

A hipótese considerada é que esses quatro fatores não são generalizáveis para todos os tipos de imagens, influenciando diretamente na qualidade da caracterização e, por conseguinte, na precisão da recuperação. Para isso, foram realizados experimentos para investigar o efeito da variação conjunta desses fatores na precisão da recuperação de imagens por conteúdo usando diferentes bases de imagens. Uma parte do conteúdo desse capítulo foi publicado em:

- PEDROSA, G.V.; TRAINA, A.J.M. Making an Image Worth a Thousand Visual Words. *XI Workshop de Visão Computacional (WVC)*, p. 1-6, Out. 2015.

3.2 A Abordagem Bag-of-Visual-Words (BoVW)

Nas últimas décadas a abordagem *Bag-of-Visual-Words* (BoVW), também denominada de *Bag-of-Features*, *Bag-of-Visual-Features* ou *Bag-of-Keypoints*, tornou-se bastante popular em várias áreas de visão computacional, como classificação de imagens, busca de vídeos, reconhecimento de textura, entre outras tarefas. Parte da sua popularidade é devido à sua simplicidade: a metodologia BoVW é baseada na representação não-ordenada de descritores locais aplicados em uma imagem e são, portanto, conceitualmente e computacionalmente mais simples do que muitos métodos alternativos. Apesar disso, ou talvez por causa disso, sistemas de recuperação que utilizam a abordagem BoVW têm apresentado um desempenho superior em vários *benchmarks* e têm conseguido avanços significativos em escalabilidade.

Na literatura pode-se encontrar diversos trabalhos que utilizam a abordagem BoVW em diferentes aplicações. Um dos pioneiros na área de recuperação de imagens foi o trabalho de (Sivic e Zisserman, 2003), que propuseram essa representação com o objetivo de detectar todas as ocorrências de um objeto nas cenas de um vídeo. Desde então outros trabalhos foram propostos para os mais diversos domínios. O trabalho de (Batista et al., 2009), por exemplo, explora a técnica BoVW com o objetivo de detectar construções em fotografias históricas; (Lopes et al., 2009) para a detecção de nudez em cenas de vídeo;

(Lazebnik et al., 2005) utiliza o método para o reconhecimento de texturas em imagens; (Ohbuchi et al., 2008) na recuperação de formas em 3D e (Yanai, 2009) na recuperação de imagens da web. Já o trabalho de (Csurka et al., 2004) utiliza a abordagem BoVW como um processo para lidar com diversos tipos de objetos e ao mesmo tempo tratar as variações de iluminação, visualização, rotação e oclusão, típicos de cenas do mundo real. Outros trabalhos, como o de (Niebles et al., 2006), exploram a representação no reconhecimento de ações em vídeos. Na área médica, (Chaudry et al., 2009) utiliza a abordagem na classificação automática de tipos de carcinoma de células renais. Mais recentemente, o trabalho de (Sun et al., 2012) é focado no problema de direitos autorais de animações, os autores buscam por fragmentos semelhantes com base nas palavras-visuais extraídas de um conjunto limitado de dados, na tentativa de recuperar fragmentos de vídeos que violam os direitos autorais.

3.2.1 Metodologia

A abordagem BoVW é uma técnica usada para modelar, de maneira robusta, as características locais de uma imagem em um único vetor-de-característica. Para isso, um *dicionário visual* é definido e cada característica local extraída da imagem é tratada como uma *palavra-visual* desse dicionário. Por meio do mapeamento dessas características locais à sua respectiva palavra-visual no dicionário, pode-se descrever a imagem em um único vetor-de-características utilizando alguma técnica de *pooling*. A técnica de *pooling* mais tradicional na literatura é o histograma de palavras-visuais que consiste em uma contagem simples da frequência com que cada palavra-visual aparece na imagem.

Figura 3.1: Metodologia clássica para representar uma imagem em *Bag-of-Visual-Words*.

O processo necessário para representar uma imagem em BoVW está esquematizado pela figura 3.1 e pode ser definido pelas seguintes fases: (i) detecção e representação das regiões-de-interesse; (ii) atribuição de cada região-de-interesse à uma palavra-visual de acordo com um dicionário (vocabulário) de palavras-visuais pré-definido; (iii) contagem da ocorrência (frequência) de cada palavra-visual contida na imagem.

Com base nessas fases, as principais decisões para a implementação do uso da abordagem BoVW em um sistema de recuperação são:

- como amostrar (detectar) as regiões-de-interesse de uma imagem;
- como descrever as regiões-de-interesse;
- como gerar o dicionário de palavras-visuais;
- como sumarizar as informações em um único vetor-de-características.

Essas quatro fases podem ser encapsuladas em duas etapas: codificação (*coding*) e sumarização (*pooling*). A etapa de codificação discretiza as características locais em palavras-visuais e a etapa de sumarização utiliza alguma abordagem que sintetiza as palavras-visuais da imagem em um único vetor-de-característica que será usado posteriormente para classificação e/ou recuperação de imagens.

A representação BoVW é flexível, uma vez que cada uma de suas fases pode ser determinada por diferentes técnicas de acordo com o domínio da aplicação. O trabalho de (Avni et al., 2011), por exemplo, apresenta um método para categorização e recuperação de imagens médicas de radiografias usando a abordagem BoVW. Para representar cada imagem, os autores inicialmente dividem a imagem em pequenas sub-regiões de tamanho 9x9 pixels, denominada *patches*, tal como ilustrado na Figura 3.2. Os valores das intensidades do nível de cinza de cada *patch* são normalizados com a finalidade de se obter média zero, ignorando-se os *patches* com intensidades uniformes. A técnica PCA é utilizada para reduzir a dimensionalidade de cada *patch* de 81 (9x9), para 7. As coordenadas do centro de cada *patch* são adicionadas, incorporando a relação espacial na caracterização de cada *patch*. A inclusão desta última informação permite analisar a disposição das palavras visuais na imagem e caracterizar, de uma maneira robusta, as informações locais.

Uma questão fundamental da abordagem BoVW se refere à decisão de qual detector de pontos-de-interesse usar, ou se deve ou não usar um detector de pontos-de-interesse, tal como avaliado por (Nowak et al., 2006). Geralmente os métodos de detecção de pontos-de-interesse mais utilizados na literatura são Harris-Affine ou Diferenças-de-Gaussianas, que é o detector utilizado pelo descritor SIFT. Porém, diversos trabalhos tem utilizado técnicas mais simples, como a detecção de pontos-de-interesse densos e aleatórios. Um detector de pontos-de-interesse denso (*Dense Sampling*) detecta

Figura 3.2: Imagem de radiografia dividida em regiões (*patches*) de tamanho 9x9 pixels. Metodologia utilizada no trabalho de (Avni et al., 2011) para detectar as regiões-de-interesse.

pontos igualmente espaçados na imagem e sem qualquer tipo de semântica envolvida nessa detecção. Além disso, outros trabalhos simplesmente utilizam a detecção aleatória desses pontos (*Random Sampling*). São estratégias bem simples, computacionalmente baratas e que tem obtido bons resultados na recuperação de imagens. Porém, não está claro ainda em quais domínios esse tipo de detecção é vantajoso.

3.2.2 Construção do Dicionário Visual

A determinação do dicionário de palavras-visuais é uma tarefa fundamental para a abordagem BoVW, pois ele é o responsável por determinar quais são as características e padrões que representam a estrutura de uma imagem. Não se pode generalizar um dicionário universal, válido para todos os tipos de aplicações, pois diferentes domínios necessitam de dicionários específicos, uma vez que a representação pode mudar. Por exemplo, os padrões representativos usando um certo descritor X é diferente das características de um outro descritor Y. Portanto, o desafio é como determinar qual é o dicionário ideal para cada tipo de aplicação de acordo com o tipo de representação utilizada para descrever as características.

Apesar da diversidade de aplicações, quase todos os trabalhos na literatura apresentam a mesma estratégia para a geração do dicionário de palavras-visuais. Essa estratégia está esquematizada na Figura 3.3 e é definida pelos seguintes passos: primeiramente um subconjunto de imagens do banco de dados é escolhido; para cada imagem, suas regiões-de-interesse são detectadas e descritas utilizando algum descritor gerando vetores-de-características; por fim, é realizado um agrupamento dos dados desse espaço de características utilizando algum algoritmo de agrupamento. O centróide de cada grupo é considerado uma palavra-visual do dicionário.

O desafio na construção de um Dicionário Visual depende de duas escolhas: o algoritmo de agrupamento e a quantidade de palavras-visuais. O algoritmo de agrupamento é um fator que interfere diretamente na qualidade e no desempenho da geração de um dicionário visual. Diferentes algoritmos de agrupamento podem levar a diferentes dicionários (Jain

Figura 3.3: Processo utilizado para a geração de um dicionário de palavras-visuais: dado um subconjunto de imagens de treinamento, suas características são extraídas e descritas. As palavras-visuais que irão compor o vocabulário do dicionário são definidas como sendo os centróides dos grupos a partir de um agrupamento realizado nesse espaço de características.

et al., 1999; Steinbach et al., 2000a), gerando resultados mais próximos das expectativas do usuário em diferentes níveis. Além disso, a *quantidade* de palavras-visuais do dicionário é outro fator que influencia diretamente na qualidade da representação da imagem. Um dicionário pequeno tem pouco poder discriminativo, uma vez que dois agrupamentos podem ser atribuídos à mesma palavra visual. Por outro lado, um vocabulário grande é pouco generalista e afeta a eficiência da abordagem. Na literatura, a quantidade de palavras-visuais é uma informação definida empiricamente e que pode variar entre diferentes bases de dados.

O agrupamento pode ser realizado, por exemplo, através do algoritmo *k-means*, que é um método de agrupamento simples e bastante utilizado na literatura. O *k-means* particiona um conjunto de pontos entre k subconjuntos disjuntos visando a minimizar a distância intra-grupos e maximizar a distância inter-grupos. Os trabalhos na literatura tendem a escolher empiricamente a quantidade de palavras visuais do dicionário representadas pelo *k-clusters* obtidos através da execução do *k-means*.

Entretanto, (Jurie e Triggs, 2005) observaram que o uso de *k-means* para a geração das palavras visuais funciona bem em imagem com texturas homogêneas, mas para imagens com objetos naturais e com regiões densas esse tipo de agrupamento não gera bons resultados para a criação de um dicionário visual. Além disso, (Santos et al., 2010) indicam que é possível substituir a abordagem de agrupamento usando o algoritmo *k-means* por uma simples seleção aleatória, sem perda estatisticamente significativa de informações, mas com uma redução radical no custo computacional.

O uso de uma técnica variante, conhecida como *bisecting k-means* (Steinbach et al., 2000a), tem demonstrado desempenho superior à técnica original no domínio da recuperação de informações textuais, mas pouco explorada na geração de palavras-visuais no domínio da recuperação de imagens. A estratégia adotada pelo algoritmo *bisecting k-means* é utilizar o agrupamento particional baseado no *k-means* sucessivamente: ele começa dividindo o conjunto de dados inicial em dois grupos (aplicando o *k-means* com k

=2), posteriormente ele seleciona um grupo resultante dessa divisão e o divide novamente em dois grupos; o algoritmo continua esse processo particional até que a quantidade k de grupos desejada seja alcançada. Essa quantidade k de grupos é uma informação fornecida antecipadamente pelo usuário.

Vários algoritmos de agrupamentos estão disponíveis na literatura, porém um problema compartilhado entre eles é a necessidade de se fornecer a quantidade de grupos *a priori*. Por exemplo, (Lazebnik et al., 2005) utiliza um vocabulário formado por 200-400 palavras visuais; (Zhang et al., 2007) adota 1.000 palavras e (Sivic e Zisserman, 2003) um vocabulário com 6.000-10.000 palavras.

Na prática, para se estimar um valor ideal de grupos, o algoritmo de agrupamento é executado múltiplas vezes variando a quantidade de grupos. Com base em uma medida de validação, o agrupamento com a melhor avaliação é escolhido (Vendramin et al., 2010). Entretanto, essa opção se torna inviável quando se lida com uma grande quantidade de dados, como é o caso deste trabalho. Logo, outras alternativas devem ser exploradas com o objetivo de prover escalabilidade, sem prejudicar a eficiência e eficácia do algoritmo.

Existem alguns métodos que permitem determinar automaticamente a quantidade de grupos. Por exemplo, o algoritmo X-means, proposto por (Pelleg e Moore, 2000) utiliza um critério de agrupamento particional utilizando uma medida conhecida como BIC (*Bayesian Information Criterion*). Entretanto, essa técnica necessita saber de antemão sobre a densidade e compacidade dos grupos, pois o algoritmo assume que todos os grupos são esféricos e com o mesmo tamanho. Isto gera *overfit* quando os grupos de dados não são esféricos.

Para um classe específica de algoritmos de agrupamento, em particular os algoritmos de agrupamento hierárquicos, não é necessário especificar a quantidade de grupos *a priori* (Steinbach et al., 2000b; Willett e El-Hamdouchi, 1989). O resultado de um algoritmo de agrupamento hierárquico pode ser graficamente visualizado como uma árvore, chamada de *dendrograma*. Esta árvore mostra graficamente o processo de fusão (*merge*) dos elementos e os *clusters* intermediários. Cortando um dendrograma em um certo nível, pode-se obter um conjunto de agrupamentos (*clusters*).

3.2.3 Fase de Codificação

Uma vez definido o dicionário de palavras-visuais, cada região-de-interesse da imagem detectada e representada por um vetor-de-característica será associada à uma palavra-visual. Em inglês, essa fase é denominada de *assignment* ou *coding*. A Figura 3.4 ajuda a ilustrar a fase de *assignment* da abordagem *Bag-of-Visual-Words*: dado que uma palavra-visual é o resultado de um agrupamento do espaço de característica, tal como discutido anteriormente, qual palavra-visual será atribuída ao triângulo amarelo?

Na literatura, a fase de codificação pode ser realizada por três diferentes abordagens:

1. *hard assignment* ou também chamada de *nearest neighbor*;
2. *multiple assignment* ou *k-nearest neighbor*;
3. *soft assignment*.

A abordagem *hard* é a mais tradicional na literatura e foi a estratégia pioneira proposta por (Sivic e Zisserman, 2003). Nessa abordagem a palavra-visual que possui a menor distância ao vetor-de-característica é determinada, e adiciona-se uma unidade ao *bin* correspondente a essa palavra-visual no histograma. No exemplo da Figura 3.4, o triângulo amarelo seria atribuído à palavra-visual w_4 , que é a mais próxima. Na abordagem *hard*, cada vetor-de-característica influencia igualmente na representação, e o histograma é normalizado de modo que cada *bin* represente a frequência relativa das palavras contidas na imagem.

Figura 3.4: Qual palavra-visual será atribuída ao triângulo amarelo?

Muitas vezes, pode ocorrer de um dado vetor-de-característica ter uma distância bem próxima entre duas palavras. No exemplo da Figura 3.4 o triângulo amarelo está bem próximo das palavras w_4 e w_5 , porém na abordagem *hard* somente a palavra-visual mais próxima w_4 seria selecionada para representar essa característica. Uma outra maneira de realizar a fase de codificação é utilizar outras abordagens como as técnicas *multiple assignment* e *soft assignment*. Essas duas estratégias foram concebidas para mitigar o impacto negativo quando um grande número de vetores-de-características de uma imagem está perto de uma fronteira de dois ou mais agrupamentos.

A abordagem *multiple assignment*, proposta por (Jegou et al., 2007), atribui ao vetor-de-características todas as k palavras-visuais mais próximas. Nesse caso, adiciona-se uma unidade ao *bin* correspondente a cada k palavra próxima. Supondo $k = 3$, o triângulo da Figura 3.4 seria atribuído às palavras w_3 , w_4 e w_5 .

A abordagem *soft assignment*, presente nos trabalhos de (Jiang et al., 2007) e (Philbin et al., 2008), também atribui mais de uma palavra-visual a um vetor-de-característica, porém ela atribui um peso de tal forma que uma palavra-visual próxima tem maior relevância do que uma palavra-visual mais distante. A técnica de codificação *soft* surgiu como uma alternativa à codificação *hard* que despreza o fato de que um

vetor-de-característica pode estar à mesma distância entre palavra-visuais diferentes. A codificação *soft* permite uma caracterização mais robusta evitando ambiguidade.

3.2.4 Fase de Sumarização

A fase de sumarização, em inglês referenciada como *pooling*, é a etapa responsável por sintetizar em um único vetor-de-característica a representação final da imagem. As três técnicas de *pooling* mais tradicionais na literatura são: *sum-pooling*, *average-pooling* e *max-pooling*. Todas elas fornecem uma representação de tamanho fixo e são baseadas na contagem da ocorrência não-ordenada das palavras-visuais no espaço da imagem.

As técnicas *sum-pooling* e *average-pooling* são semelhantes: cada uma gera um histograma de palavras-visuais, porém a diferença entre elas está no fato de que *sum-pooling* representa a soma das ocorrências de cada palavra-visual na imagem, enquanto *average-pooling* se refere à média, ou seja, um histograma normalizado. Já a técnica de *max-pooling*, quando usada com a técnica de codificação *hard*, gera um vetor binário, indicando a presença ou ausência de uma palavra-visual na imagem. Quando *max-pooling* é usado com a codificação *soft*, gera um histograma indicando a máxima ocorrência de cada palavra-visual na imagem.

Para ilustrar, a Figura 3.5 mostra o exemplo da aplicação dessas três técnicas de *pooling* em uma imagem que tem 5 pontos-de-interesse (p1,...,p5), um dicionário composto por 3 palavras-visuais (A,B,C) e usando 2 técnicas de codificação diferentes (*hard* e *soft*). As três técnicas geram representação diferentes e que depende da codificação usada.

Pontos	Palavras-visuais			Pontos	Palavras-visuais		
	A	B	C		A	B	C
p1	0,5	0,2	0,3	p1	1	0	0
p2	0,7	0	0,3	p2	1	0	0
p3	0,3	0,2	0,5	p3	0	0	1
p4	0,5	0	0,5	p4	1	0	0
p5	0,1	0,4	0,5	p5	0	0	1
avg	0,42	0,16	0,42	avg	0,6	0	0,4
sum	2,1	0,8	2,1	sum	3	0	2
max	0,7	0,4	0,5	max	1	0	1

(a)
(b)

Figura 3.5: Exemplo da aplicação das técnicas de *pooling* (*avg*, *max* e *sum*) em uma imagem com 5 pontos e 3 palavras-visuais usando: (a) *soft assignment* e (b) *hard assignment*.

Outras técnicas de *pooling* mais sofisticadas codificam informação de distância na representação final. Por exemplo, a ideia da técnica BossaNova proposta por (Avila et al., 2013) é computar um histograma de distâncias entre cada característica local extraída da imagem e suas respectivas palavras-visuais. A técnica VLAD (*Vector of*

Locally Aggregated Descriptors) de (Jegou et al., 2012) acumula, para cada palavra-visual, a diferença entre a distância de cada ponto à sua respectiva palavra-visual. Esses métodos de sumarização permitem caracterizar a distribuição dos vetores-de-características locais aos seus respectivos centros, que correspondem às palavras-visuais.

O grande problema com essas técnicas é a perda da informação espacial das palavras-visuais na imagem. Alguns trabalhos recentes mostram que a inclusão da localização espacial das palavras-visuais é uma característica bastante importante em uma técnica de *pooling* para aumentar o poder discriminativo da abordagem.

3.3 Avaliações Experimentais

Para implementar a abordagem BoVW em um sistema CBIR quatro grandes questões devem ser consideradas para se obter uma caracterização semântica mais próxima à expectativa do usuário:

1. Qual a melhor maneira de detectar e representar as regiões-de-interesse de uma imagem?
2. Como gerar um dicionário eficiente e eficaz?
3. Como codificar e sumarizar de maneira eficiente cada região-de-interesse?
4. Como medir a similaridade?

Diversas técnicas têm sido propostas na literatura para atender esses requisitos, porém continua sendo um campo de pesquisa com oportunidades para avanços e diversas lacunas que podem e devem ser exploradas para aumentar ainda mais a eficiência da abordagem em um sistema CBIR.

Nesta seção são discutidos e avaliados em detalhes quatro importantes aspectos da abordagem BoVW: o tamanho do dicionário visual, o detector das regiões-de-interesse, a técnica de codificação e a função de distância utilizada para medir a similaridade entre imagens. A proposta é investigar esses quatro parâmetros em diferentes bases e como eles afetam em conjunto a precisão dos resultados na recuperação de imagens por conteúdo.

3.3.1 Bases Avaliadas

Os experimentos foram realizados usando cinco diferentes bases de imagens, todas elas de domínio público e disponíveis na literatura:

- Base 1: é a base Corel1000¹, que consiste em uma base de dados formada por cenas naturais bem complexas e cheias de detalhes, composta por 1.000 imagens divididas em 10 classes. A Figura 3.6a ilustra um exemplo de cada classe;

¹base disponível em: <http://wang.ist.psu.edu/docs/related/>

- Base 2: é formada por imagens de texturas utilizada no trabalho de (Lazebnik et al., 2005), composta por imagens de superfícies de materiais, tais como madeira, mármore e pele sob diferentes pontos de vista, escalas e condições de iluminação. O conjunto consiste em 1.000 imagens de tamanho 640x480 pixels, divididos em 40 amostras de 25 diferentes classes. A Figura 3.6b mostra alguns exemplos de algumas classes;
- Base 3: base composta por 5.042 imagens biomédicas de 32 categorias diferentes, tais como Raio-X, CT, MRI, etc. Ela é um subconjunto da base do ImageCLEFmed 2007 (Henning Müller, 2008). A Figura 3.6c mostra alguns exemplos dessa base;
- Base 4: base 15-Scenes (Lazebnik et al., 2006), composta por 4.485 imagens de cenas naturais separadas em 15 categorias. Cada categoria é composta por 210 a 410 imagens, e o tamanho médio das imagens é de 300x250 pixels. A Figura 3.6d mostra um exemplo de cada classe;
- Base 5: base de imagens da Oxford Flowers (Nilsback e Zisserman, 2006), composta por 1.360 imagens de 17 espécies diferentes de flores(80 imagens por categorias), com variações de escala, iluminação e oclusões parciais. A Figura 3.6e mostra alguns exemplos de flores de espécies diferentes dessa classe.

Em conjunto, todas essas bases abrangem uma grande diversidade de tipos de imagens, cada uma com características específicas e aplicações variadas. O objetivo aqui é explorar as situações em que cada técnica possui um desempenho superior e analisar como os resultados podem mudar de um domínio para outro.

3.3.2 Abordagens Avaliadas

Quatro diferentes detectores de pontos-de-interesse foram avaliados nos experimentos:

- Harris-Affine (Harris);
- Diferença-de-Gaussianas (DoG);
- Pontos Aleatórios (*Random Sampling*);
- Denso (*Dense Sampling*).

A Figura 3.7 mostra o resultado da aplicação desses detectores em algumas imagens.

Para descrever os pontos foi utilizada a proposta do descritor SIFT, que é baseada na magnitude e gradiente de uma região 16x16 ao redor de cada ponto-de-interesse detectado. O agrupamento utilizado para gerar o dicionário visual foi obtido usando a estratégia do algoritmo *bisecting k-means*. As codificações utilizadas foram *hard* e *soft* e a técnica de sumarização *average-pooling*.

Figura 3.6: Alguns exemplos de imagens de cada base utilizada nos experimentos.

3.3.3 Resultados Obtidos

Os gráficos da Figura 3.8 mostram os valores de mAP (*mean Average Precision*) obtidos em cada base avaliada e usando diferentes tamanhos de dicionário, entre 310 e 2500. Em geral, na literatura os melhores resultados são obtidos usando dicionários grandes, acima de 1.000 palavras-visuais. Porém, pelos experimentos realizados nota-se que, para a maioria das bases, aumentando o tamanho do dicionário os valores de mAP não aumentam expressivamente, em média apenas 0.05% . Em alguns casos, por exemplo para a base 4 (15-scenes), o aumento do tamanho do dicionário fez reduzir a precisão da recuperação. Isso mostra que dicionários muito grandes não são muito vantajosos, o pouco ganho na precisão (e até a perda) não compensa o custo computacional exigido por um vetor-de-característica de alta dimensionalidade.

Além disso, a grande maioria dos trabalhos propostos na literatura defendem a codificação *soft* sobre a *hard*, por ser menos ambígua e gerar uma caracterização com menor erro de quantização. Porém, outro fato importante observado nos experimentos foi

Figura 3.7: Resultado da aplicação dos detectores de pontos-de-interesse utilizados nos experimentos.

que, em todas as bases, os melhores resultados foram obtidos através da codificação *hard*, porém a diferença de precisão entre as codificações *hard* e *soft* são pequenas em algumas bases (2,4,5) e pode chegar até 0,2% em outras (bases 1,3). Apenas para a base 3, a codificação *soft* utilizando o detector DoG obteve resultado superior à codificação *hard*.

O fato mais importante observado foi que o detector de pontos-de-interesse tem um forte impacto na precisão da recuperação. Os melhores resultados foram obtidos usando os detectores Harris e *Dense*. Esse último, por sua vez, obteve os melhores resultados em três bases (1,3,4). O detector DoG por sua vez, que é a técnica utilizada pelo descritor SIFT, gerou resultados semelhantes ao detector baseado em pontos-aleatórios (*Random*). Inclusive para duas bases (1 e 3), o detector *Random* obteve uma acurácia quase semelhante ao detector *Dense*. Isso mostra que, para algumas bases, não é necessário utilizar detectores de pontos-de-interesse “sofisticados”, ou seja, técnicas que possuam alguma semântica envolvida na detecção dos pontos. Pelos experimentos notou-se que as bases que não necessitam de detectores sofisticados são aquelas formadas por imagens de cenas naturais e médicas. Para bases complexas, como texturas e flores, o detector Harris se apresentou como o mais adequado.

Foi analisada também a influência da função distância na precisão final da recuperação de imagens. Para isso, foram realizados testes usando seis diferentes funções de distâncias, L_1 , L_2 , χ^2 , Canberra, Cosseno e Intersecção de Histogramas. As equações dessas funções

Figura 3.8: Valores de mAP (mean Average Precision) obtidos usando diferentes: tamanho de dicionários, detectores de pontos-de-interesse e técnicas de codificação.

foram definidas no Capítulo 2. A Tabela 3.1 apresenta os valores de mAP obtidos para as bases avaliadas usando um dicionário formado por 310 palavras-visuais, codificação *hard* e usando o detector de pontos-de-interesse *Dense*.

Nota-se que a função de distância escolhida também exerce influência na acurácia da recuperação de imagens. Para a maioria das bases (1,4,5) as funções de distâncias L_1 e Intersecção de Histogramas apresentaram os melhores resultados. Apenas para as bases 2 e 3, a função de distância Canberra apresentou resultado superior, porém não muito melhor que as funções L_1 e Intersecção de Histogramas. Desse modo, a distância L_1 se apresentou a mais vantajosa, pois é a distância que tem menor custo computacional dentre as avaliadas.

Tabela 3.1: mAP valores obtidos usando diferentes funções de distâncias.

Base	Função de Distância					
	L_1	L_2	χ^2	Canberra	Cosseno	Intersecção
Base 1 (Corel)	0,503	0,436	0,419	0,465	0,472	0,503
Base 2 (Textura)	0,268	0,230	0,245	0,321	0,243	0,266
Base 3 (Médica)	0,712	0,641	0,672	0,715	0,680	0,712
Base 4 (15-Scenes)	0,568	0,504	0,385	0,493	0,519	0,568
Base 5 (Flores)	0,118	0,113	0,114	0,114	0,113	0,118

3.4 Considerações Finais

Neste capítulo foi apresentada em detalhes a abordagem *Bag-of-Visual-Words* (BoVW). Tal abordagem, apesar de sua popularidade na área, tem muitos parâmetros que afetam o seu desempenho, por exemplo o método de geração do dicionário, o tamanho do dicionário, o tipo de detector utilizado, a sumarização final da imagem, etc. Todas essas fases podem ser implementadas usando diferentes técnicas. O fato é que o bom desempenho final depende fortemente dessa escolha, o que não é uma tarefa fácil, pois como avaliado pelos experimentos, não se pode generalizar um conjunto de técnicas que atenda todos os diferentes conjuntos de imagens.

A modelagem BoVW simples e tradicional é pouco descritiva e apresenta grandes oportunidades para o avanço de pesquisas com o objetivo de desenvolver um ferramental que possa caracterizar imagens de maneira efetiva, usando informações que possam quantificar o melhor possível a similaridade perceptual entre as imagens. Nos próximos capítulos serão apresentadas algumas propostas desenvolvidas neste trabalho de doutorado que permitiram ampliar a representação semântica da abordagem BoVW e estendê-la para outros tipos de caracterização, por exemplo para formas (silhuetas) de objetos. Como consequência, os trabalhos propostos nesta tese possibilitaram aumentar a acurácia e o

poder de descrição da abordagem baseada em palavras-visuais na tarefa de recuperação de imagens em sistemas CBIR.

Abordagens Estendidas para o Modelo

Bag-of-Visual-Words

4.1 Considerações Iniciais

A modelagem *Bag-of-Visual-Words* (BoVW) pode ser definida, sucintamente, em duas fases: a primeira fase (*coding*) é responsável por codificar as características locais extraídas da imagem em palavras-visuais através de um dicionário-visual; a segunda fase (*pooling*) sumariza o conjunto das palavras-visuais em um único vetor-de-características, que será usado nos processos de classificação e/ou recuperação da imagem. A técnica tradicional de *pooling* utilizada pela abordagem BoVW é a simples contagem da frequência de cada palavra-visual contida na imagem, gerando um *histograma de palavras-visuais*, sem considerar qualquer relação de ordem entre elas, daí o nome *Bag* (saco). A principal vantagem de se utilizar um histograma como a representação final da imagem é o baixo custo computacional na sua geração/construção, uma vez que a complexidade computacional é linear em relação ao número de pontos (palavras) detectados na imagem.

No entanto, a capacidade descritiva de um histograma é extremamente limitada. Esta limitação está relacionada ao fato de que uma única palavra-visual contém pouca informação semântica, o que não é o ideal quando se deseja codificar de maneira efetiva as características visuais de uma imagem. Esta situação pode ser exemplificada por uma analogia com a representação de documentos textuais: palavras isoladas apresentam uma capacidade muito limitada de descrição, por exemplo, as palavras “inteligência” ou “artificial” não necessariamente contribuiriam para classificar um documento textual como sendo do tópico “ciências”, porém quando estas duas palavras aparecem juntas no texto essa probabilidade aumenta. Este exemplo mostra que, se as palavras são organizadas com

uma relação de co-ocorrência, a representação semântica de um texto aumenta e então pode-se obter uma descrição de alto-nível que seja similar à percepção humana. Por esta razão, a codificação espacial das palavras-visuais no domínio das imagens deve ser levado em consideração para aumentar a capacidade descritiva de uma imagem. Porém, o desafio é: como codificar o conjunto das palavras-visuais presentes em uma imagem de uma forma eficaz e eficiente?

Nos últimos anos, alguns trabalhos têm sido propostos para integrar a falta da informação espacial no modelo BoVW, tal como será vista na seção 4.2. Estas técnicas se diferem na estratégia utilizada na análise espacial que pode ser, por exemplo, baseada em regiões ou em distâncias entre pontos. Porém, ao se criar uma abordagem de codificação espacial algumas propriedades são desejáveis, tais como: invariância às transformações geométricas, eficiência computacional e representação compacta. A maioria dos métodos propostos na literatura são de alta complexidade computacional e não são invariantes a transformações geométricas. Desse modo, ainda existem oportunidades para o desenvolvimento de técnicas que permitam incluir a informação espacial de uma maneira mais eficaz e eficiente.

Neste capítulo são apresentadas três novas técnicas de *pooling* que agregam a informação espacial das palavras-visuais na caracterização final da imagem, são elas: (i) *Weighted Histogram* (WH), (ii) *Bunch-of-2-grams* (Bo2g) e (iii) *Global Spatial Arrangement* (GSA). Cada uma dessas técnicas tem sua particularidade de descrição e são propostas complementares. A primeira técnica (WH) gera um *histograma ponderado*, analisando a frequência das palavras-visuais contidas em uma imagem de acordo com a região espacial onde cada palavra se situa. A segunda técnica (Bo2g) codifica as relações de co-ocorrências entre palavras, agregando as palavras-visuais em *frases-visuais*. A terceira técnica (GSA) é mais global e analisa a *disposição das palavras* em todo o espaço da imagem, ou seja, a técnica GSA permite estimar em qual região da imagem (topo, esquerda, direita ou abaixo) uma palavra-visual tende a se situar na imagem. Essas três técnicas geram descrições semanticamente mais enriquecidas sobre o conjunto das palavras-visuais contido em uma imagem, permitindo caracterizações mais robustas quando comparadas a um simples histograma. As técnicas apresentadas nesse capítulo foram publicadas em:

1. PEDROSA, G.V.; TRAINA, A.J.M. Compact and Discriminative Approach for Encoding Spatial-Relationship of Visual Words. In: *ACM/SIGAPP Symposium On Applied Computing (SAC)*, Salamanca, Spain, p. 92-95, 2015.
2. PEDROSA, G.V.; RAHMAN, M.; ANTANI, S. K. ; DEMNER-FUSHMAN, D.; LONG, L. R.; TRAINA, A.J.M. Integrating Visual Words as Bunch of n-Grams for Effective Biomedical Image Classification. In: *IEEE Winter Conference on*

Applications of Computer Vision (WACV), Steamboat Springs, USA, p. 431-436, 2014.

3. PEDROSA, G.V.; TRAINA, A.J.M. From Bag-of-Visual-Words to Bag-of-Visual-Phrases Using n-Grams. In: *Conference on Graphics, Patterns and Images (SIBGRAPI)*, Arequipa Peru, p. 304-311, 2013.
4. PEDROSA, G.V.; REZENDE, S.O.; TRAINA, A.J.M. Reducing the dimensionality of the SIFT descriptor and increasing its effectiveness and efficiency in image retrieval via bag-of-features. In: *Brazilian Symposium on Multimedia and the Web (WebMedia)*, São Paulo, Brazil, p. 139-142, 2012.

4.2 Trabalhos Correlatos

A principal limitação do uso do histograma como uma abordagem de *pooling* na metodologia *Bag-of-Visual-Words* é o fato que ele codifica apenas as estatísticas de primeira ordem do conjunto de palavras-visuais na imagem, ignorando assim as estatísticas espaciais, tais como co-ocorrências das palavras-visuais, que são informações importantes na análise da similaridade entre imagens. Um dos primeiros trabalhos que discutiram esta importância é a obra de (Sivic e Zisserman, 2003) que formalizou o modelo BoVW como um descritor para imagens. Os autores apresentaram uma analogia com um procedimento muito comum utilizado pelo Google: aqueles documentos em que as palavras da consulta aparecem juntas no texto tem prioridade no *ranking* de resposta. Essa analogia é especialmente relevante para a consulta de objetos usando uma sub-parte da imagem. Desde então, a informação espacial de palavras visuais tem sido tema de pesquisas.

Nesta seção é apresentada uma extensa revisão bibliográfica dos trabalhos que tratam o problema da codificação espacial na abordagem BoVW. A Figura 4.1 mostra uma linha do tempo com as obras que serão discutidas. As técnicas foram agrupadas em duas abordagens diferentes: baseadas em região e baseadas em distância. Essa divisão foi realizada de acordo com a estratégia utilizada pelos trabalhos na análise espacial. Os trabalhos baseados em região dividem a imagem em regiões, enquanto que os trabalhos baseados em distância codificam a informação espacial diretamente sobre os *pixels* da imagem usando algum critério espacial, tal como a distância entre as palavras-visuais.

4.2.1 Codificação Espacial Baseada em Distância

Analisar a vizinhança é a estratégia-chave usada pelos trabalhos baseados em distância. O trabalho de (Sivic et al., 2005), por exemplo, procura aumentar a descrição semântica incrementando o dicionário visual com “doublets”, que são pares de palavras-visuais que co-ocorrem dentro de uma vizinhança espacial. Esses *doublets* são formados por todos os

Figura 4.1: Linha-do-tempo dos trabalhos correlatos.

pares de palavras-visuais que estão dentro dos 5-vizinhos mais próximos de cada palavra. Apesar da relevância do método, os autores não apresentaram o ganho discriminatório dessa técnica na recuperação de imagens.

Seguindo a estratégia de *doublets*, o modelo Bi-gram de (Lan et al., 2007) usa um dicionário formado por “primitivas”, tais como *singlets*, *doublets* e *triplets*. O modelo Bi-gram consiste no cálculo das probabilidades de transição entre primitivas correspondentes vizinhas. Dois *doublets* são considerados vizinhos se eles compartilham uma palavra-visual, enquanto *triplets* são considerados vizinhos se possuem duas palavras-visuais em comum. Para eliminar a ambiguidade de correspondência, todos os *doublets* com palavras repetidas são removidos. Experimentalmente, os autores demonstraram a capacidade do modelo Bi-gram em incrementar o reconhecimento de objetos que são afetados por transformações não-rígidas, como o espelhamento.

A abordagem PIWAH, proposta por (Khan et al., 2012), também analisa pares de palavras-visuais para codificar a disposição espacial. A ideia dessa técnica é calcular a orientação (ângulo) dos segmentos formados por pares de palavras-visuais semelhantes. A descrição final é dada por um histograma dos ângulos formados pelos pares. A técnica apresenta bom desempenho se utilizada com um tamanho de dicionário pequeno, porém não é uma abordagem robusta à rotação das imagens.

A abordagem Δ -TSR de (Hoàng et al., 2010) usa relações triangulares, explorando a co-ocorrência de pontos triplos. A técnica codifica na descrição final a geometria triangular e as orientações relativas de todas as combinações triplas de todos os pontos. O método, apesar de eficaz, tem alta complexidade computacional para construir e podar os triângulos. Além de co-ocorrências, o método GVP de (Zhang et al., 2011) usa um conjunto de palavras (frases) de comprimento n , definido como todas as n palavras-visuais em uma determinada disposição espacial (*layout*). Palavras diferentes e *layouts* diferentes definem frases diferentes. O método GVP muitas vezes gera um bom *top 10* de imagens recuperadas, no entanto frases grandes tem um desempenho ruim em geral. Essa ideia de frases também é encontrada no trabalho de (Yuan et al., 2007), com a abordagem “Visual Phrase Lexicon”, que define uma frase visual como sendo um conjunto de palavras-visuais que co-ocorrem espacialmente formando um padrão e utilizando um método de teste

da razão de probabilidade que avalia a significância das palavra-visuais, o que é uma estratégia computacionalmente cara.

Outros trabalhos têm sido propostos para a codificação de informação espacial usando vizinhança e distribuição de distância. O trabalho em (Savarese et al., 2006) usa a ideia de “correlogramas” de palavras-visuais para modelar as correlações espaciais entre as palavras-visuais. Um correlograma expressa como a co-ocorrência de um par de palavra-visuais muda com a distância, e esta abordagem pode capturar tanto informação local quanto global, basta apenas mudar o raio de abrangência. Esta estratégia é bem adequada quando se deseja representar um objeto inteiro ou apenas uma parte dele. Porém, é uma técnica dependente do valor do raio de abrangência na construção dos correlogramas.

A codificação da distribuição de distância também é utilizada pelo método “Binned Log-Polar” de (Zhang e Mayo, 2010). Essa abordagem usa o centro da imagem como o ponto de referência e gera uma grade log-polar. Em seguida, calcula-se um histograma de palavras-visuais em cada célula dessa grade. A representação final é uma concatenação de todos os histogramas, criando um grande vetor-de-característica único, porém não-invariante à rotação. Essa estratégia de grade log-polar também é usado no método BoSVW de (Bolovinou et al., 2013). Porém, para lidar de maneira eficiente com a alta dimensionalidade do vetor final, o método BoSVM emprega um agrupamento baseado em K-means para gerar uma representação espacial mais compacta.

Visando uma representação mais refinada e específica para cada classe da base de dados, o trabalho de (Feng et al., 2011) define uma norma ponderada ℓ_p -norm de distribuição espacial específica para cada classe da base de dados. O método tem como objetivo estimar uma função de *pooling* ponderada que codifica a informação geométrica da distribuição das palavras para cada classe específica do conjunto de dados. A principal desvantagem da técnica é que ela não consegue lidar com um conjunto de dados genérico (por exemplo, a base PASCAL VOC), especialmente quando a base de dados contém grandes variações intra-classe. Isto é devido ao fato de que a abordagem está firmemente adaptada às posições das palavras. Para bases de dados maiores e mais gerais, o “Spatial Codebook” de (Mbanya et al., 2011) é mais adequado, uma vez que ele usa uma estratégia muito simples: incluir no vetor-de-característica final, a localização de cada ponto na imagem detectada, com a desvantagem da não-invariância a rotação.

Abordagens baseadas em grafos estão emergindo como modelos alternativos e mais robustos para analisar a informação espacial. A abordagem BoVG de (Silva et al., 2013) codifica as relações espaciais das palavras-visuais como um conjunto de grafos conectados. Esta informação é usada para a construção de um segundo vocabulário: um dicionário baseado em grafos. Este dicionário permite a construção de um histograma de grafos conexos. No entanto, o tamanho ideal do dicionário é uma questão que necessita de aprimoramento e ajustes.

4.2.2 Codificação Espacial Baseada em Regiões

Rumo a uma representação visual semanticamente superior, os trabalhos baseados em região utilizam uma estratégia diferente para codificar a informação espacial de palavras-visuais em comparação com as abordagens baseadas em distância. Ao invés de analisar a vizinhança ou a distribuição de distância, os trabalhos baseados em região dividem a imagem em regiões. Por exemplo, a técnica “Bundled Features” de (Wu et al., 2009) codifica a informação espacial por meio da concatenação de histogramas de palavras-visuais em regiões da imagens segmentadas através do método MSER. Porém, esta abordagem é computacionalmente demorada, uma vez que a verificação espacial é realizada durante o processo de recuperação.

Usando características de atenção visual seletiva, o objetivo da abordagem “Spatial Weighting” de (Marszaek e Schmid, 2006) é dar pesos menores para as palavras-visuais pertencentes à região de fundo da imagem. Desse modo, as palavras-visuais pertencentes ao objeto tem um peso maior na representação final. Seguindo essa mesma estratégia, o modelo FISM de (Soares et al., 2012) computa dois histogramas: um que representa as palavras-visuais pertencentes à região do objeto e outro que representa as palavras-visuais que pertencem ao fundo da cena da imagem. Para isso eles exploram modelos de saliências e técnicas *fuzzy* para representar a incerteza de classificar a posição de uma palavra como sendo pertencente ao objeto ou ao fundo da imagem.

Recentemente, a técnica WSA de (Penatti et al., 2014) codifica a informação espacial das palavras-visuais analisando o seu comportamento em relação às outras palavras-visuais detectadas na imagem. A técnica WSA subdivide a imagem em quadrantes, utilizando cada ponto detectado como a origem desse quadrante. Intuitivamente, esta estratégia analisa o arranjo das palavras em relação aos outros pontos da imagem, desse modo, pode-se estimar se uma palavra-visual tende a ser inferior, à direita, ou cercada por outras palavras. A principal desvantagem da técnica é a não-invariância a rotação. Isso significa que, uma imagem rotacionada terá uma representação diferente em comparação com sua versão original (não rotacionada).

Usando uma combinação de adjacência espacial com segmentação baseada em super-pixels, o modelo SABOF de (Tao et al., 2014) calcula a frequência de cada palavra-visual, construindo um histograma com a informação espacial. Para tornar o modelo mais robusto, um histograma é calculado para cada super-pixel gerado usando vários valores de super-pixels.

A abordagem mais popular na literatura para codificar a informação espacial de palavras visuais é a abordagem “Spatial Pyramid” (SP) proposta por (Lazebnik et al., 2006). Esta técnica subdivide a imagem em células hierárquicas e calcula uma *Bag-of-Visual-Words* para cada célula, concatenando todos os histogramas gerados no final, tal como ilustrado na Figura 4.2. Os resultados desta abordagem mostraram ser

Figura 4.2: A estratégia da técnica *Spatial Pyramid* (SP) é dividir a imagem em células hierárquicas e computar uma BoVW para cada região. Essa abordagem além de não ser invariante à rotação, gera um vetor com alta dimensionalidade.

consideravelmente superiores em desempenho na classificação de imagens. No entanto, as desvantagens da abordagem SP são: a não-invariância à transformação geométrica e o grande vetor-de-características gerado, tornando impraticável o seu uso em grande bases de dados. O modelo “Spatial-BoF” de (Cao et al., 2010) lida com a não-invariância da abordagem SP, alterando a ordem dos *bins* dos histogramas. O histograma de cada palavra-visual é re-ordenado a partir do *bin* com a maior frequência. No entanto, esta re-ordenação nem sempre corresponde à verdadeira transformação da imagem.

4.3 Métodos Propostos: Agregando a Informação Espacial

A tradicional modelagem *Bag-of-Visual-Words* baseada em histograma é semanticamente pouco descritiva. Portanto, o objetivo geral dos métodos que serão propostos nesta seção é ampliar a caracterização semântica da imagem para uma efetiva recuperação de imagens baseada em conteúdo. Para isso, a ideia é tirar vantagem da distribuição espacial das palavras-visuais na representação final da imagem.

As técnicas que serão apresentadas fornecem *três* valiosas informações sobre o conjunto das palavras-visuais contidas em uma imagem:

- *Weighted Histogram* (WH): frequência ponderada de palavras-visuais;
- *Bunch-of-2-Grams* (Bo2g): co-ocorrência entre as palavras-visuais;
- *Global Spatial Arrangement* (GSA): disposição global das palavras-visuais no espaço da imagem.

Nas próximas subseções serão apresentadas em detalhes cada proposta, após uma definição formal do problema.

4.3.1 Definição Formal do Problema

Seja $P = \{p_1, p_2, \dots, p_N\}$, o conjunto não-ordenado de N pontos-de-interesse detectados em uma imagem, tal que cada ponto $p_i \in \mathbb{R}^D$, em que D é a dimensionalidade do vetor-de-característica associado ao ponto p_i . Esse conjunto P pode ser detectado e descrito usando qualquer descritor, por exemplo o descritor SIFT, desse modo $D = 128$.

Supõe-se que um *dicionário-visual* com K palavras-visuais foi previamente obtido utilizando alguma técnica de agrupamento, por exemplo *K-means*, e que tal dicionário seja representado pelo conjunto $C = \{c_1, \dots, c_K\}$, em que $c_i \in \mathbb{R}^D$.

A construção da modelagem *Bag-of-Visual-Words* pode ser decomposta em duas fases sequenciais: codificação (*coding*) e sumarização (*pooling*). A fase de codificação pode ser modelada por uma função $f : \mathbb{R}^D \rightarrow \mathbb{R}^K$, em que:

$$p_i \rightarrow f(p_i) = \{\alpha_1^i, \alpha_2^i, \dots, \alpha_K^i\} \quad (4.1)$$

em que α_j^i pode ser vista como uma função de ativação, associando o ponto p_i à j -ésima palavra-visual do dicionário. Essa ativação pode ser *hard* ou *soft*. A codificação clássica da abordagem BoVW é baseada na codificação *hard* em que cada ponto p_i é atribuído à apenas uma palavra-visual, ou seja:

$$\alpha_j^i = \begin{cases} 1, & \text{se } j = \underset{j \in \{1, \dots, K\}}{\operatorname{argmin}} \|p_i - c_j\|^2 \\ 0, & \text{caso contrário} \end{cases} \quad (4.2)$$

Ou então, pode-se utilizar uma codificação *soft* atribuindo valores diferentes para cada palavra-visual do dicionário. A tradicional técnica *soft* considera a distância de cada ponto à palavra-visual da seguinte maneira:

$$\alpha_j^i = \frac{\exp(-\beta \|p_i - c_j\|_2^2)}{\sum_{j'=1}^K \exp(-\beta \|p_i - c_{j'}\|_2^2)} \quad (4.3)$$

em que β é um parâmetro que controla a suavidade da codificação.

A etapa de *pooling* ocorre após a etapa de codificação, e a técnica de *pooling* utilizada pela abordagem BoVW é dada por uma função $g : \mathbb{R}^{N \times K} \rightarrow \mathbb{R}^K$, tal que:

$$g(\{\alpha_j^i\}) = h = [h_1, h_2, \dots, h_K], \quad i = \{1, \dots, N\} \text{ e } j = \{1, \dots, K\} \quad (4.4)$$

em que:

$$h_j = \sum_{i=1}^N \alpha_j^i, \quad j = \{1, \dots, K\} \quad (4.5)$$

Em outras palavras, a técnica tradicional de *pooling* dada pela Eq. 4.5 gera um histograma de palavras-visuais. Esse histograma é semanticamente pouco discriminatório, pois duas imagens muito diferentes podem ter as mesmas frequências de palavras. Logo, o objetivo aqui é *agregar a informação da disposição espacial das palavras-visuais* na representação final da imagem, visando ampliar a caracterização semântica da imagem para uma efetiva recuperação de imagens baseada em conteúdo. A proposta desse trabalho para a integração dessa codificação espacial na representação final da imagem é dada por três técnicas diferentes e complementares que serão formalizadas a seguir.

4.3.2 Distribuição Ponderada de Frequências

Um aspecto não capturado pela abordagem BoVW é a *relevância visual* de cada palavra-visual (ponto) na representação final da imagem. Por exemplo, uma palavra-visual contida em uma região homogênea da imagem tem menos “importância visual” do que uma palavra-visual contida em uma região não-homogênea, como bordas. Por esse motivo, uma maneira de enriquecer a caracterização e a análise da similaridade entre imagens é considerar que as palavras-visuais detectadas tem diferentes contribuições para se distinguir efetivamente duas imagens e, por esse motivo, poderiam ser subestimadas/descartadas na representação final. A partir dessa observação, a ideia proposta neste trabalho é dar uma relevância diferente (menor) para aquelas palavras-visuais que estão em *regiões homogêneas*, já que visualmente aquelas palavras que estão em *regiões não-homogêneas* tem uma relevância maior.

A estratégia proposta neste trabalho para capturar a relevância visual de um ponto em uma imagem é analisar se este ponto está em uma região em que vários de seus vizinhos possuem a mesma palavra-visual. Se sim, então este ponto está numa região homogênea. A ideia é considerar de maneira diferente, na representação final da imagem, aqueles pontos que possuem muitos vizinhos próximos com a mesma palavra-visual, pois a motivação é que estes pontos têm uma relevância visual menor e leva a um excesso de redundância na caracterização da imagem.

Formalmente, o *conjunto dos pontos vizinhos* a um dado ponto p_i dentro de um raio de abrangência r , é definido por:

$$\phi_{vizinho}(p_i) = \{p_j \in P \mid dist(p_i, p_j) \leq r\} \quad (4.6)$$

em que $dist(p_i, p_j)$ é a distância do ponto p_i ao ponto p_j . O *conjunto dos pontos vizinhos similares* ao ponto p_i são aqueles que possuem a mesma palavra-visual, ou seja:

$$\phi_{similar}(p_i) = \{p_j \in \phi_{vizinho}(p_i) \mid f(p_j) = f(p_i)\} \quad (4.7)$$

A partir da Equação 4.7, pode-se estimar se um ponto p_i detectado em uma imagem está (ou não) em uma região homogênea da imagem analisando a quantidade de pontos vizinhos que compartilham a mesma palavra-visual: se todos os vizinhos tem a mesma palavra-visual, então o ponto detectado está em uma região homogênea.

Já que todo ponto p_i está associado à uma palavra-visual, então ao invés de computar um histograma “uniforme” de palavra-visuais, tal como faz a abordagem BoVW tradicional dada pela Eq. 4.5, neste trabalho é proposto um Histograma Ponderado, (*Weighted Histogram*), dando pesos diferentes àqueles pontos contidos em regiões diferentes na imagem.

Definição 4.3.1 *O Histograma Ponderado (Weighted Histogram - WH) de uma imagem descrita em palavras-visuais, é definido por:*

$$h_j = \sum_{i=1}^N w(p_i) \alpha_j^i, \quad j = \{1, \dots, K\} \quad (4.8)$$

em que:

$$w(p_i) = 1 - \frac{|\phi_{similar}(p_i)|}{|\phi_{vizinho}(p_i)|} \quad (4.9)$$

Na Equação 4.9, as notações $|\phi_{similar}(p_i)|$ e $|\phi_{vizinho}(p_i)|$ correspondem a quantidade de elementos em cada conjunto. Esta equação gera um histograma ponderado de frequências, em que cada palavra-visual contribui de modo diferente na construção do histograma. Essa contribuição é dada pela Eq. 4.9, que atribui um peso menor para aquelas palavras-visuais que possuem muitos vizinhos com a mesma palavra-visual. Se $|\phi_{vizinhos}(p_i)| = |\phi_{similar}(p_i)|$, ou seja, se todos os vizinhos são similares, então $w(p_i) = 0$. Quando $|\phi_{vizinhos}(p_i)| \rightarrow |\phi_{similar}(p_i)|$, ou seja, quanto mais vizinhos forem mais similares, menor é o peso atribuído e, portanto, menor é a contribuição da palavra-visual associada a este ponto na representação final da imagem.

A Figura 4.3 mostra visualmente um exemplo da aplicação da técnica de ponderação proposta usando dois diferentes detectores de pontos-de-interesse: denso e Diferenças-de-Gaussianas (DoG). Quanto mais escuro é o ponto na imagem, menos importante ele é considerado na representação proposta. Quando o raio de abrangência $r = 0$, todos os pontos tem o mesmo “peso”. Quando r é máximo, os pontos importantes são aqueles únicos, ou seja, aqueles que correspondem à uma palavra-visual única em toda a imagem. Quando $0 < r < \text{máximo}$, então os pontos mais importantes são aqueles que não tem nenhum ponto vizinho com a mesma palavra-visual.

Figura 4.3: Exemplo mostrando a proposta de ponderação de pontos. Quanto mais escuro, menos importante é o ponto e menor é a contribuição da palavra-visual associada a este ponto na representação final da imagem.

4.3.3 Co-ocorrências Baseada em Bag-of-Visual-Phrases

Ao invés de representar as imagens em palavras-visuais isoladas, uma descrição mais robusta pode ser obtida, por exemplo, agrupando palavras e gerando uma nova representação que poderia ser chamada de *Bag-of-Visual-Phrases*. A Figura 4.4 mostra um esquema ilustrativo dessa ideia comparando as técnicas *Bag-of-Visual-Words* e *Bag-of-Visual-Phrases*. A representação *Bag-of-Visual-Phrases* seria mais robusta, pois conseguiria capturar a relação espacial entre as palavras e, desse modo, ao invés de trabalhar com um dicionário de palavras-visuais, a ideia seria contemplar a relação espacial das palavras-visuais usando um dicionário-de-frases.

No entanto representar uma imagem em frases-visuais não é uma tarefa trivial, já que não é possível determinar a *ordem* e o *conjunto* de palavras que determinariam uma frase-visual. A possibilidade de combinação de todas as frases que poderiam ser formadas com K palavras-visuais seria de ordem exponencial K^K . Para contornar esse problema, poderia ser utilizada uma alternativa de caracterização que utilizasse frases com tamanhos definidos. Na área textual, essas frases com tamanhos fixos são chamados de n -grams, em que n é o tamanho de palavras que compõem a frase. Porém, ainda assim, a possibilidade seria exponencial com relação o tamanho de n . Por esse motivo, a motivação neste trabalho foi trabalhar com frases baseadas em 2-grams (bi-grams), pois reduz o espaço de possibilidades de frases para K^2 .

Figura 4.4: Metodologia *Bag-of-Visual-Words* e *Bag-of-Visual-Phrases*.

Definição 4.3.2 Neste trabalho, o conjunto de 2-grams que pode ser detectado em uma imagem é definido por:

$$2\text{-grams}(P) = \{(f(p_i), f(p_j)) \mid \text{dist}(p_i, p_j) < r, \forall i, j \in \{1, \dots, N\}\} \quad (4.10)$$

A Eq. 4.10 define que um bi-grama é formado por todos os pares de pontos que estão dentro de uma região de vizinhança, dada por um raio de abrangência r . A partir da Definição 4.3.2, pode-se definir uma nova modelagem, denominada aqui de *Bag-of-2-grams*, que é definida como a frequência de todos os possíveis bi-grams extraídos na imagem.

Definição 4.3.3 Dado um conjunto de 2-grams extraído de uma imagem, a modelagem *Bag-of-2-grams* é definida por:

$$\text{Bag-of-2-grams}(2\text{-grams}(P)) = \begin{bmatrix} b_{11} & \dots & b_{1K} \\ \vdots & \ddots & \vdots \\ b_{K1} & \dots & b_{KK} \end{bmatrix} \quad (4.11)$$

em que,

$$b_{ij} = \{|B_{ij}| \subset 2\text{-grams}(P) \mid (i, j) \in B_{ij}\} \quad (4.12)$$

A Figura 4.5 ilustra um exemplo comparando a representação *Bag-of-Visual-Words* e *Bag-of-2-grams*: dado um conjunto de palavras-visuais, os bi-grams são gerados definindo

um raio de abrangência em cada ponto da imagem e, então, a representação final é dada computando-se a ocorrência (frequência) dos bi-grams gerados

Figura 4.5: Exemplo mostrando a construção do modelo *Bag-of-2-grams*.

A principal desvantagem da abordagem *Bag-of-2-grams* é a dimensionalidade do vetor-de-características gerado. O tamanho desse vetor é K^2 , em que K é a quantidade de palavras-visuais do dicionário visual. Se o tamanho do dicionário visual é muito grande, esta abordagem torna-se inviável nos processos de recuperação e/ou classificação da imagem.

Para reduzir ainda mais a dimensionalidade final do vetor-de-característica, e sem perder o poder descritivo, neste trabalho é proposta uma estratégia para agrupar conjuntos de bi-grams “similares”, gerando um vetor mais compacto. Para isso, a estratégia proposta é medir a entropia de *Shannon* sobre conjuntos de bi-grams, gerando uma nova modelagem denominada *Bunch-of-2-grams*.

Definição 4.3.4 Dado o conjunto de *Bag-of-2-grams* gerado a partir da definição 4.3.3, a modelagem *Bunch-of-2-grams* (Bo2g) é definida como:

$$Bo2g(Bag-of-2-grams) = [b'_1, b'_2, \dots, b'_K] \quad (4.13)$$

em que,

$$b'_i = - \sum_{j=1}^K b_{ij} \cdot \ln(b_{ij}) \quad (4.14)$$

A proposta do descritor *Bunch-of-2-grams* (Bo2g) é estimar, em um único valor, a quantidade de bi-grams que tenham no mínimo uma palavra-visual em comum. Essa estratégia permite reduzir um vetor-de-características com dimensão K^2 para um vetor com apenas K valores.

4.3.4 Arranjo Espacial Global

As duas técnicas propostas até o momento são abordagens locais, ou seja, analisam a disposição espacial das palavras-visuais em relação à uma vizinhança limitada ao redor de cada palavra. A terceira técnica de *pooling* proposta neste trabalho, denominada de *Global Spatial Arrangement* (GSA), analisa a disposição espacial *global* das palavras-visuais no espaço da imagem. A estratégia é dividir a imagem em um quadrante $\mathcal{Q} = \{Q1, Q2, Q3, Q4\}$, usando cada palavra-visual como o centro desse quadrante, e então contar, de maneira incremental, a ocorrência de cada palavra-visual em cada região desse quadrante. Intuitivamente, essa contagem irá estimar as posições das palavras em relação aos outros pontos detectados na imagem, ou seja, será possível estimar onde uma palavra-visual mais ocorre na imagem: acima, abaixo, esquerda ou direita.

Formalmente, o espaço de imagem é dividido da seguinte maneira: para cada ponto p_i detectado na imagem, a imagem é dividida em um quadrante $\mathcal{Q} = (Q1, Q2, Q3, Q4)$, colocando o ponto p_i na origem deste quadrante e sendo Q1 a região superior esquerda, Q2 a região superior direita, Q3 a região inferior esquerda e Q4 a região inferior direita, tal como ilustrado pela Figura 4.6. Então, para qualquer outro ponto p_j , incrementa-se a contagem da palavra-visual associada a este ponto p_j na sua posição correspondente no quadrante. Por exemplo, supondo que A é a palavra-visual associada ao ponto p_j , e p_j está no quadrante superior direito em relação ao ponto p_i , então o contador de Q2 é incrementado. Esse procedimento é realizado para todos os outros pontos. Depois que todos os pontos forem analisados em relação à p_i , a origem do quadrante vai para o próximo ponto p_{i+1} , e a contagem em relação a p_{i+1} continua. Quando todos os pontos já tiverem sido origens do quadrante, a contagem termina.

Figura 4.6: Um quadrante é definido utilizando cada ponto p_i detectado na imagem, então incrementa-se a ocorrência das palavras-visuais em cada região desse quadrante.

A Figura 4.7 mostra um exemplo ilustrando passo-a-passo a aplicação da estratégia utilizada. Esta técnica permite estimar onde uma palavra-visual tende a se localizar em

relação às outras palavras-visuais. Se uma palavra-visual tem valores diferentes de zero apenas na região inferior direita do quadrante, então isso significa que esta palavra aparece sempre na posição inferior direita em relação aos outros pontos detectados na imagem. No exemplo da Figura 4.7, a palavra B aparece apenas na parte superior em relação aos outros pontos, pois os contadores da parte inferior são 0. Por outro lado, os contadores da palavra C estão quase igualmente distribuídos, isso significa que C é uma palavra que está envolvida pelos outros pontos na imagem.

Figura 4.7: Exemplo da estratégia utilizada para codificar o arranjo espacial das palavras visuais. Os círculos vermelhos são os pontos detectados com a sua palavra-visual associada. Coloca-se a origem de um quadrante em cada ponto detectado e, então, conta-se, de maneira incremental, o número de palavras-visuais que aparecem em cada região do quadrante. Os valores finais do contador são exibidas na imagem inferior direita.

Invariância a Rotação

No entanto, a estratégia apresentada anteriormente irá gerar uma representação diferente se a imagem é rotacionada. Para solucionar este problema, neste trabalho é utilizada a informação da direção do gradiente de cada palavra-visual para definir cada região do quadrante de uma maneira mais robusta a transformações de rotação. A direção do gradiente é invariante à rotação, porque este sempre aponta na direção da maior variação de intensidade, tal como ilustrado pela Figura 4.8. Devido a este fato, esta informação é utilizada para definir qual região do quadrante (Q1, Q2, Q3 ou Q4) um ponto $p_j = (x_j, y_j)$ pertencerá considerando a direção do gradiente do ponto de origem $p_i = (x_i, y_i)$.

O gradiente é uma mudança de direção na intensidade dos pontos de uma imagem. Matematicamente, o gradiente de um ponto p_i é um vetor $G(p_i) = \{G_x(p_i), G_y(p_i)\}$, com

Figura 4.8: O gradiente sempre aponta para a região onde ocorre a maior variação de intensidade.

suas componentes representadas pelas derivadas em relação à direção horizontal G_x e vertical G_y . Usando apenas o sinal das componentes G_x e G_y , neste trabalho é proposta uma estratégia invariante à rotação, dada pela definição a seguir.

Definição 4.3.5 *Seja $p_i = (x_i, y_i)$ o ponto-origem de um quadrante $\mathcal{Q} = (Q1, Q2, Q3$ e $Q4)$, em que $Q1$ é a região superior esquerda, $Q2$ a região superior direita, $Q3$ a região inferior esquerda e $Q4$ a região inferior direita, e $G(p_i) = \{G_x(p_i), G_y(p_i)\}$ as componentes do gradiente do ponto p_i . Para definir, de maneira invariante à rotação, qual região desse quadrante um ponto $p_j = (x_j, y_j)$ irá percenter, são definidas as seguintes condições:*

Se $G_x(p_i) > 0$ e $G_y(p_i) > 0$, então

se $x_j < x_i$, e $y_j < y_i$, então $p_j \rightarrow Q1$

se $x_j < x_i$, e $y_j > y_i$, então $p_j \rightarrow Q2$

se $x_j > x_i$, e $y_j < y_i$, então $p_j \rightarrow Q3$

se $x_j > x_i$, e $y_j > y_i$, então $p_j \rightarrow Q4$

Se $G_x(p_i) > 0$ e $G_y(p_i) < 0$, então

se $x_j < x_i$, e $y_j > y_i$, então $p_j \rightarrow Q1$

se $x_j < x_i$, e $y_j < y_i$, então $p_j \rightarrow Q3$

se $x_j > x_i$, e $y_j > y_i$, então $p_j \rightarrow Q2$

se $x_j > x_i$, e $y_j < y_i$, então $p_j \rightarrow Q4$

Se $G_x(p_i) < 0$ e $G_y(p_i) > 0$, então

se $x_j > x_i$, e $y_j < y_i$, então $p_j \rightarrow Q1$

se $x_j > x_i$, e $y_j > y_i$, então $p_j \rightarrow Q3$

se $x_j < x_i$, e $y_j < y_i$, então $p_j \rightarrow Q2$

se $x_j < x_i$, e $y_j > y_i$, então $p_j \rightarrow Q4$

Se $G_x(p_i) < 0$ e $G_y(p_i) < 0$, então

se $x_j > x_i$, e $y_j > y_i$, então $p_j \rightarrow Q1$

se $x_j > x_i$, e $y_j < y_i$, então $p_j \rightarrow Q2$

se $x_j < x_i$, e $y_j > y_i$, então $p_j \rightarrow Q3$

se $x_j < x_i$, e $y_j < y_i$, então $p_j \rightarrow Q4$

Usando as condições da Definição 4.3.5, a representação será robusta à rotação, pois cada região do quadrante é definida pelo sinal do gradiente do ponto de origem. A Figura

Figura 4.9: A direção do gradiente é usada para definir cada região do quadrante. Essa estratégia torna a representação invariante à rotação.

4.9 ilustra a Definição 4.3.5, em que as setas em vermelho representam as direções x e y do gradiente do ponto de origem. Se o $G_x(p_i)$ é negativo, então o gradiente aponta para cima, caso contrário para baixo. Se $G_y(p_i)$ é positivo, então o gradiente aponta à direita da origem, senão para esquerda.

Representação Compacta

A estratégia apresentada gera 4 valores para cada palavra-visual da imagem, estimando a disposição de cada palavra-visual nas quatro regiões do quadrante. Considerando um dicionário visual composto por K palavras-visuais, o tamanho do vetor-de-característica final seria $4K$. Para reduzir a dimensionalidade da representação final, neste trabalho é proposta a ideia de considerar cada região Q1, Q2, Q3 e Q4 em apenas dois valores de acordo com a posição de cada região do quadrante: $top = (Q1+Q2)$, $down = (Q3+Q4)$, $left = (Q1+Q3)$, $right = (Q2+Q4)$, tal como ilustrado pela Figura 4.10. A representação final é dada pela relação $top/down$, $left/right$, definida a seguir.

Figura 4.10: A representação final proposta agrupa cada região do quadrante em relação às suas posições: top , $down$, $left$ e $right$.

Definição 4.3.6 A representação final da técnica GSA é dada por 2 valores ($S_{t/d}, S_{l/r}$), considerando a relação entre os grupos top-down e left-right, definidas como:

$$S_{t/d} = \frac{top}{top + down} \quad (4.15)$$

$$S_{l/r} = \frac{left}{left + right} \quad (4.16)$$

Usando a definição acima, cada palavra-visual será associada a apenas dois valores, indicando a disposição espacial de cada palavra nas regiões acima (*top*) e esquerda (*left*) das palavras detectadas na imagem. Desse modo, a dimensionalidade do vetor-de-característica final diminui de $4K$ para $2K$.

4.4 Avaliações Experimentais

Diversas avaliações experimentais foram realizadas usando cinco diferentes bases de imagens. As bases analisadas são as mesmas utilizadas nas avaliações experimentais do Capítulo 3, são elas: Corel1000, Textura, Flowers, 15-Scenes e Médicas.

Para descrever os pontos foi utilizada a proposta do descritor SIFT, com codificação *hard* e a técnica de sumarização *average-pooling*. O agrupamento utilizado para gerar o dicionário visual foi obtido usando a estratégia do algoritmo *bisecting k-means*, e o tamanho do dicionário foi definido em $K = 310$ palavras-visuais.

As três abordagens propostas (WH, Bo2g e GSA) foram comparadas com duas outras técnicas da literatura: *Spatial Pyramid* (SP) proposta por (Lazebnik et al., 2006) e *Words Spatial Arrangement* (WSA) proposta por (Penatti et al., 2014), bem como com a tradicional técnica BoVW. A Tabela 4.1 mostra o tamanho do vetor-de-características utilizado em cada técnica avaliada, bem como o tempo médio para extrair as características e o tempo gasto na análise da similaridade. O descritores BoVW e SP são os mais rápidos para extrair as características, porém o descritor SP é o mais lento de todos os descritores avaliados para analisar a similaridade entre duas imagens, pois tem o maior vetor-de-características. Os descritores propostos (WH, Bo2g e GSA) tem pequenos vetores-de-características e, por esse motivo, são rápidos para computar a similaridade entre duas imagens.

Influência dos parâmetros na precisão da recuperação

O descritor *Weighted Histogram* (WH) necessita de um parâmetro r que indica o raio de abrangência necessário para computar a quantidade de vizinhos similares a partir de um ponto. Uma maneira heurística para estimar o valor de r é utilizar a média das distâncias dos pontos detectados na imagem. Para isso, foram realizados diversos experimentos usando cinco diferentes valores de $r = \{0, \mu - \sigma, \mu, \mu + \sigma, max\}$. Quando $r = 0$, tem-se

Tabela 4.1: Tempo e espaço utilizado pelo descritores avaliados.

Descritor	Tamanho do Vetor	Tempo Médio [μs]	
		Extração	Análise da Similaridade
BoVW	$K = 310$	3,28	1,27
SP	$K + 20K = 6.510$	72,44	4,85
WSA	$K + 4K = 1.550$	9.636,03	2,60
GSA	$K + 2K = 930$	9.756,70	1,89
WH	$K = 310$	526,46	1,28
Bo2g	$K + K = 620$	8.946,99	1,57

um histograma sem ponderação; os valores μ e σ representam, respectivamente, a média e o desvio padrão das distâncias dos pontos detectados na imagem, e max é o valor máximo do raio que engloba toda a imagem.

Os gráficos da Figura 4.11 mostram os valores de mAP (*mean Average Precision*) obtidos para cada base usando diferentes valores de r e dois detectores de pontos-de-interesse diferentes: denso e Harris. Quando r está próximo da média das distâncias entre os pontos, a precisão da recuperação aumenta e fica acima da precisão quando $r = 0$. Portanto, neste trabalho, uma maneira automática de estimar r é usar a média das distâncias entre os pontos.

4.4.1 Resultados Obtidos

Para uma análise mais robusta com relação a transformações geométricas (rotação), foi realizada uma extensão de cada base: para cada imagem foi gerada uma cópia rotacionada 180 graus. Dessa maneira, a quantidade de imagens em cada base foi duplicada. A base Corel1000 passa a ter 2.000 imagens, a base de Texturas 2.000 imagens, a base Flower 2.720, a base 15-Scenes 8.970 e a base Médica 10.084 imagens.

Os gráficos da Figura 4.12 mostram os valores de mAP obtidos em cada base usando os descritores avaliados. Os valores $mAP1$ e $mAP2$ se referem, respectivamente, aos valores de mAP obtidos usando a base original e a base estendida/duplicada com as imagens rotacionadas. Os descritores SP e WSA apresentam bons desempenhos nas bases originais, no entanto esses dois descritores não são invariantes à rotação, por esse motivo perdem desempenho quando a base contém imagens similares porém rotacionadas. Essa situação não é observada nos descritores propostos (WH, Bo2g e GSA), todos eles tem ótimo desempenho tanto na base original, quanto na base estendida.

A Tabela 4.2 mostra o ganho dos descritores avaliados em cada base com relação à tradicional técnica BoVW. O descritor SP tem os maiores ganhos na bases originais ($mAP1$), porém perde precisão nas bases estendidas ($mAP2$), pois o descritor não é

Figura 4.11: Valores de mAP (mean Average Precision) obtidos usando diferentes valores de r , em que μ e σ representam a média e o desvio padrão das distância entre os pontos detectados na imagem.

Figura 4.12: Valores de mAP (*mean Average Precision*) obtidos utilizando cada descritor. Os valores de $mAP1$ se referem às bases originais e $mAP2$ se referem às bases estendidas com imagens rotacionadas.

invariante à rotação. Os descritores propostos (WH, Bo2g e GSA) não perdem precisão

em nenhuma das bases. O descritor WH, por exemplo, tem os melhores ganhos na base Médica. Já o descritor Bo2g apresentou os melhores ganhos na base de imagens de Texturas. O descritor GSA apresenta o melhor desempenho na base Corel formada por imagens de cenas naturais.

Tabela 4.2: Ganho obtido em cada descritor avaliado quando comparado à técnica tradicional BoVW baseada em histogramas de palavras-visuais. Os valores de $mAP1$ se referem às bases originais e $mAP2$ se referem às bases estendidas com imagens rotacionadas.

Base		Descritor				
		SP	WSA	GSA	WH	Bo2g
Corel	$mAP1$	★ 12,9%	10,2%	11,2%	8,4%	5,0%
	$mAP2$	-3,7%	-7,0%	★ 9,5%	8,7%	3,5%
Textura	$mAP1$	5,7%	3,0%	4,5%	4,2%	★ 8,4%
	$mAP2$	0,5%	0,3%	3,4%	5,9%	★ 10,1%
Flower	$mAP1$	★ 18,2%	12,7%	10,5%	5,8%	8,9%
	$mAP2$	-8,2%	-6,8%	13,5%	10,7%	★ 14,8%
15-Scenes	$mAP1$	★ 7,4%	3,9%	1,9%	7,3%	2,6%
	$mAP2$	-0,6%	-1,3%	1,9%	★ 5,2%	2,7%
Médica	$mAP1$	6,8%	5,0%	4,1%	★ 7,7%	6,4%
	$mAP2$	-1,8%	-3,1%	4,5%	★ 8,6%	6,8%

4.5 Considerações Finais

Neste capítulo foram apresentados três novos descritores que estendem a abordagem BoVW, permitindo incluir a disposição espacial das palavras-visuais na caracterização final da imagem.

O primeiro descritor proposto, chamado de *Weighted Histogram* (WH), gera um histograma de palavras-visuais ponderado. Na tradicional técnica BoVW, cada palavra-visual detectada na imagem contribui com uma unidade no *bin* correspondente a esta palavra no histograma final. A estratégia do descritor WH é considerar cada palavra-visual detectada na imagem de modo diferente na geração do histograma. A ideia é que uma palavra-visual que esteja rodeada por várias palavras-visuais semelhantes tenha um peso menor no histograma final da imagem. Essa maneira de caracterização permite enriquecer a análise da similaridade entre duas imagens, uma vez que é possível destacar palavras-visuais que estejam em regiões visualmente relevantes na imagem. A maior vantagem do descritor WH é que ele não altera a dimensionalidade do vetor-de-características com relação à técnica

BoVW, permitindo aumentar a eficácia da modelagem BoVW sem afetar a eficiência na análise da similaridade.

O segundo descritor é motivado pela ideia de frases-visuais baseada em n-gramas, chamado de *Bunch-of-2-grams* (Bo2g). A ideia é codificar a interação entre pares de palavras-visuais vizinhas, gerando uma nova representação chamada de 2-gramas. Porém, a quantidade possível de combinação entre palavras-visuais é exponencial com relação à quantidade de palavras do dicionário-visual. Para contornar tal problema, o descritor Bo2g calcula a entropia de pares de palavras-visuais formadas com uma palavra em comum. Essa estratégia de caracterização permite codificar a interação entre as palavras de uma maneira mais compacta.

Por fim, o terceiro descritor é mais global e analisa a disposição das palavras-visuais em todo o espaço da imagem, chamado de *Global Spatial Arrangement* (GSA). A ideia desse descritor é analisar a ocorrência das palavras-visuais em quadrantes, usando cada palavra como o centro desse quadrante. A representação final é invariante à rotação, pois utiliza a direção do gradiente para definir cada região do quadrante. Além disso, a representação final utiliza apenas dois valores, pois o descritor agrupa as informações de cada quadrante em relação a posição de cada região do quadrante: *top*, *down*, *left*, *right*.

Comparados com a abordagem BoVW tradicional, a principal vantagem dos descritores propostos é permitir incorporar a disposição espacial das palavras-visuais na caracterização final da imagem, de uma maneira compacta e robusta a transformações geométricas (rotação). Essa informação extra permite enriquecer a análise da similaridade entre duas imagens e aumentar a precisão na recuperação de imagens em até 14%.

Descrição de Formas em *Bag-of-Salience-Points*

5.1 Considerações Iniciais

A percepção visual de uma imagem pelo ser humano é uma tarefa subjetiva e por isso é difícil representá-la por meio de uma única característica. Cor, forma e/ou textura são as tradicionais informações mais utilizadas para descrever uma imagem e dentre essas três, a identificação, reconhecimento e recuperação de formas têm utilidade em diversas aplicações, por exemplo, no reconhecimento de caracteres alfanuméricos, na detecção e reconhecimento de pessoas em sistemas de segurança, no rastreamento de objetos em vídeos, etc.

A análise de formas pode ser feita usando diferentes abordagens, e na literatura encontram-se diversos descritores para tal fim (Arulmozhi e Abirami, 2014). Porém, a análise e caracterização de formas ainda é um problema da área de visão computacional que apresenta oportunidades para novas abordagens e possíveis ajustes em técnicas já propostas. O desafio é que um descritor de formas deve ser robusto a transformações geométricas, além de extrair apenas características discriminativas e compactas para, assim, obter um sistema de recuperação eficaz e eficiente.

Neste capítulo é apresentado um novo descritor para caracterizar e analisar a similaridade de formas explorando o uso de pontos de saliência. Este descritor tem uma forte inspiração cognitiva, já que saliências concentram as informações visuais mais importantes e são, portanto, características preciosas que auxiliam a distinguir perceptualmente de maneira efetiva duas formas. O novo descritor proposto modela os pontos-de-saliência de uma forma em *Bag-of-Salience-Points* usando um *Dicionário de Curvaturas*. Essa

nova abordagem, que é uma analogia com a modelagem *Bag-of-Visual-Words*, permite caracterizar e medir a similaridade entre formas usando funções de distâncias computacionalmente rápidas.

Avaliações experimentais foram realizadas usando diferentes bases de imagens e o desempenho da técnica proposta foi comparado com vários outros descritores da literatura. O novo descritor tem alto desempenho na recuperação de formas, baixo custo computacional e um poderoso poder descritivo, apresentando uma valiosa contribuição para o estado-da-arte na área. Parte do conteúdo relacionado a esse capítulo foi publicado em:

- PEDROSA, G.V.; TRAINA, A.J.M.; BARCELOS, C.A.Z. Shape Description Based on Bag-of-Saliency-Points. In: *ACM/SIGAPP Symposium On Applied Computing (SAC)*, Salamanca, Spain, p. 74-79, 2015.

5.2 Contextualização

5.2.1 Desafios e Taxonomia

Apesar da importância das formas na representação da imagem, o desenvolvimento de um descritor de formas não é uma tarefa fácil e trivial na área de visão computacional. Para uma máquina, a forma é um dos aspectos mais difíceis de se detectar, devido à dificuldade de segmentar os objetos de interesse presentes na imagem. Entende-se por segmentação o processo de extração de objetos relevantes em uma imagem, o que não é uma tarefa trivial, pois diferentes objetos requerem diferentes algoritmos ou parâmetros, o que torna complexa a tarefa de segmentação.

Os descritores de formas exigem, muitas vezes, que as imagens sejam previamente segmentadas e que, após o processo de segmentação, os objetos de interesse na imagem se encontrem separados do fundo e de outros objetos irrelevantes, podendo ser utilizados diretamente pelos descritores de formas. Por este motivo, *geralmente as propostas de descritores de formas na literatura já supõem que os objetos nas imagens foram segmentadas em um processo anterior*. Esta estratégia também é utilizada neste trabalho.

Outra razão que torna o desenvolvimento de um descritor de formas difícil para a área de visão computacional é o fato de que um objeto pode apresentar uma rica variabilidade de formas e diferentes objetos podem apresentar formas com uma alta semelhança visual. Esse problema pode ser observado pelas formas da Figura 5.1. Cada imagem representa a silhueta de objetos pertencentes a diferentes classes, mas com alta similaridade visual. Um descritor deve ser capaz de representar essas sutis diferenças e, ao mesmo tempo, proporcionar robustez às variações geométricas que podem ocorrer com as formas, tais como a rotação, translação e escala. Além disso, outras propriedades importantes são desejáveis para um descritor de forma, tais como: representação compacta, tempo

de processamento rápido para extrair as características e para medir a similaridade. Além da dificuldade da similaridade interclasses, há também dificuldades decorrentes da variabilidade intraclasses.

Figura 5.1: Formas de objetos diferentes porém com alta similaridade visual.

Vários métodos para reconhecer e descrever formas foram e continuam sendo desenvolvidos, uma revisão sistemática de vários métodos propostos pode ser encontrada em (Arulmozhi e Abirami, 2014). Em suma, esses métodos podem ser classificados em dois grandes grupos: métodos baseados em contorno e métodos baseados em região. Os métodos baseados em contornos apenas analisam o contorno da forma, enquanto os métodos baseados em região analisam a forma (o objeto) como um todo. A Figura 5.2 ilustra esta classificação.

Figura 5.2: Classificação das técnicas de descrição e representação de formas.

As abordagens baseadas em contorno são mais populares na literatura, entretanto elas podem sofrer efeitos negativos na presença de ruído no contorno. Por esse motivo, a abordagem baseada em região pode ser mais robusta na recuperação de formas por considerar a informação de toda a forma ao invés de considerar apenas a informação do contorno.

Cada uma das abordagens (contorno e região) são ainda subdivididos em global e estrutural, como visto na Figura 5.2. A diferença entre essas duas abordagens está no fato de que na abordagem global não se divide a forma em sub-partes, enquanto que na abordagem estrutural a forma é dividida em segmentos e cada método determina um tipo diferente de primitiva. Entretanto, a abordagem estrutural é mais complexa computacionalmente, especialmente para o cálculo de distância, além disso é mais sensível a ruído.

Exemplos de descritores baseados em contorno com abordagem global são os Descritores de Fourier (Dalitz et al., 2013), e com abordagem estrutural tem-se o *Curvature Scale Space* (CSS), proposto em (Mokhtarian e Bober, 2011). Os descritores baseados em Momentos Invariantes (Liao e Pawlak, 1996) são exemplos de métodos regionais com abordagem global, e analisar uma forma por meio do seu fecho-convexo é uma maneira de descrever uma forma em regiões com uma abordagem estrutural.

5.2.2 Motivação

Quando os seres humanos olham para uma cena complexa, sem qualquer conhecimento prévio, existe pouco tempo para processar toda a informação visual disponível, portanto sua atenção é atraída apenas para uma pequena parte dessa informação (a mais relevante). Essa parte relevante então é processada com um maior nível de detalhes e as regiões menos importantes podem (ou não) ser processadas posteriormente. Essa maneira de compreender a semântica de uma imagem conduz a um tratamento sequencial de diferentes partes da cena visual de uma imagem, isto torna possível a análise de cenas complexas em tempo real com recursos limitados de processamento. Este processo de seleção e ordenação é chamado na literatura de *atenção seletiva* (Zhang e Lin, March 2013).

Uma maneira de analisar a similaridade entre formas, que é mais próxima à análise utilizada pelo ser humano, é analisar os seus pontos de saliências, tal como explorado nos trabalhos de (Pedrosa et al., 2013) e (Torres e Falcão, 2007). Um ponto saliente (ou ponto de canto) é um ponto com elevado grau de curvatura ao longo do contorno da forma, tal como ilustrado pelos exemplos na Figura 5.3. Computacionalmente esses pontos salientes são muito valiosos para a descrição de formas, porque estão relacionados com as regiões visualmente mais importantes da forma, o que pode levar à uma representação compacta e mais próxima a percepção humana. Porém, apesar dos pontos salientes serem invariantes à rotação e translação, outros aspectos contribuem para a qualidade dessa abordagem na análise de formas, tais como: a sensibilidade do algoritmo de detecção de saliências em formas com contorno ruidoso, o método de representar os pontos de saliências e, principalmente, a medida de similaridade usada para calcular a distância entre duas formas descritas por seus pontos de saliências.

Figura 5.3: Pontos de saliências de algumas formas.

A principal desvantagem de se utilizar saliências como características é o custo computacional para calcular a similaridade entre duas formas. Este problema está

relacionado ao fato de que duas formas podem ter quantidades diferentes de pontos salientes, sendo assim, não se pode usar funções de distância tradicionais, como a distância Euclidiana, para medir a similaridade, pois os vetores não terão dimensionalidade iguais.

Neste trabalho é proposto um novo descritor de formas via identificação e caracterização de saliências, referenciado como *Bag-of-Saliency-Points* (BoSP). O método proposto é inspirado na abordagem *Bag-of-Visual-Words*, em que a ideia é atribuir uma palavra-visual à cada ponto de saliência detectado no contorno da forma. A representação final é um único vetor-de-características com tamanho fixo, que corresponde ao histograma de palavras-visuais. Essa abordagem de caracterização permite analisar a similaridade usando funções de distâncias computacionalmente mais rápidas, como a distância Euclidiana ou City-block, como será discutido na próxima subseção.

5.3 Método Proposto

O descritor de formas proposto BoSP (*Bag-of-Saliency-Points*) está esquematizado na Figura 5.4 e consiste nos seguintes passos:

1. Os pontos salientes do contorno da forma são detectados usando um método evolutivo baseado em um filtro anisotrópico;
2. Depois de detectados, estes pontos salientes são representados em um vetor-de-características, que representa os valores de curvatura em multi-escala das saliências no contorno da forma;
3. Atribui-se uma palavra-visual à cada ponto detectado e caracterizado com base em um Dicionário de Curvaturas;
4. A representação final é dada por um Histograma Global de palavras-visuais combinado com a proposta de Histogramas Espaciais que codificam a distribuição das distâncias das palavras-visuais no espaço forma em relação ao centróide.

As próximas subseções descrevem em detalhes cada um desses passos. A combinação de todas as propostas que serão apresentadas permitem descrever formas de maneira compacta e robusta, com a vantagem de se poder utilizar funções de distâncias simples e rápidas para medir a similaridade entre as formas representadas pelos seus pontos de saliências.

5.3.1 Detecção das Saliências

Para extrair os pontos salientes do contorno da forma, foi utilizado o detector proposto em (Pedrosa e Barcelos, 2010). A ideia é utilizar um filtro anisotrópico para eliminar

Figura 5.4: Esquema do descritor BoSP.

os pontos ruidosos do contorno, que não são saliências, mas também apresentam alto valor de curvatura. Essa técnica de suavização é não-linear, uma vez que ela efetua uma suavização mais fraca em determinados pontos e mais forte em outros. Isso permite que um ponto de saliência verdadeiro não seja eliminado. A principal vantagem dessa nova abordagem de detecção é que ela é robusta a ruídos e evita que o usuário tenha que definir vários parâmetros.

A metodologia do detector é a seguinte: seja $B = \{p_1, \dots, p_m\}$ o conjunto formado pelos pontos do contorno de uma forma, ordenados percorrendo o contorno no sentido horário (ou anti-horário), em que $p_i = (x_i, y_i)$, para $i = 1, \dots, m$, em que m é o número de pontos do contorno e x_i e y_i representam as coordenadas do ponto p_i na imagem. Para detectar os pontos de saliências, três fases são necessárias:

1. Primeira fase consiste em calcular o conjunto $\tilde{B} = \{\Psi_t(p_1), \dots, \Psi_t(p_m)\}$, em que $\Psi_t(p_i)_{i=1\dots m}$ representa o valor da curvatura do ponto p_i . Dado um ponto $p_i = (x_i, y_i)$ e seus t -vizinhos $p_{i+t} = (x_{i+t}, y_{i+t})$ e $p_{i-t} = (x_{i-t}, y_{i-t})$, o cálculo discreto da curvatura do ponto p_i é dado por:

$$\Psi_t(p_i) = \frac{(x_{i+t} - x_{i-t})(y_{i+t} - 2y_i + y_{i-t}) - (y_{i+t} - y_{i-t})(x_{i+t} - 2x_i + x_{i-t})}{((x_{i+t} - x_{i-t})^2 + (y_{i+t} - y_{i-t})^2)^{3/2}} \quad (5.1)$$

2. No segundo passo, com o objetivo de reduzir os altos valores de curvatura dos pontos ruidosos, é aplicada uma suavização no conjunto $\tilde{B}$ usando um filtro de suavização anisotrópica. Matematicamente, dada uma função $I \subset \mathbb{R}^p \rightarrow \mathbb{R}$, $p = 1, 2$ ou 3 , o filtro anisotrópico é definido como:

$$\begin{aligned} u_t &= \text{div}(g(|\nabla B|)\nabla B) \\ u_0 &= I \end{aligned} \tag{5.2}$$

em que I é o sinal ruidoso, g é uma função tal que $g(0) = 1$, $g(s) \geq 0$, e $g(s) \rightarrow 0$ quando $s \rightarrow \infty$. A ideia é que se $|\nabla u|$ é grande, então a difusão será pequena e, portanto, a localização de um ponto onde há mudança brusca de valor será preservada. Portanto, a função g permite uma difusão localizada adaptativa.

3. Depois de suavizado o contorno, a terceira fase consiste em localizar em $\tilde{B}$ os pontos de máximo e mínimo locais que serão os pontos de saliência.

Após esses três passos, cada forma será representada pelos seus pontos de saliência: $\tilde{B} = \{s_1, \dots, s_n\}$, em que n é a quantidade de saliências da forma.

5.3.2 Representação das Saliências

Após a detecção dos pontos salientes, é preciso descrever cada saliência com um valor representativo. Neste trabalho, é proposta a estratégia de análise do segmento de contorno onde a saliência está localizada, usando uma representação multi-escala da curvatura.

Dado um ponto p_i do contorno da forma, para calcular a curvatura deste ponto, podemos usar a Equação 5.1 apresentada na subseção anterior. Esta equação é uma implementação discreta de uma equação de curvatura contínua, que precisa de um parâmetro t para representar a vizinhança no cálculo da derivada. Variando o valor de t podemos obter uma representação em multi-escala do contorno da forma. Além disso, para tornar a representação invariante à escala, o valor de t deve ser definido com base no número de pontos do contorno da forma, dessa maneira se o tamanho da forma é aumentado ou diminuído, a representação não é afetada por esta transformação.

A representação em multi-escala de cada ponto de saliência é obtida utilizando-se diferentes valores de t . Neste trabalho, foram utilizados 5 valores diferentes de $t = \{\frac{m}{\alpha}, \frac{m}{2\alpha}, \frac{m}{3\alpha}, \frac{m}{4\alpha}, \frac{m}{5\alpha}\}$, em que m é o número de pontos do contorno da forma, α é uma constante tal que $5\alpha < m$.

Formalmente, seja $\tilde{B} = \{s_1, \dots, s_n\}$ os pontos de saliência extraídos da forma usando a técnica descrita na seção anterior. Cada ponto de saliência s_i será representada por um vetor com dimensão 5, que representa os valores de curvatura calculado usando 5 valores diferentes de t , ou seja, $s_i = [\Psi_{\frac{m}{\alpha}}(s_i), \Psi_{\frac{m}{2\alpha}}(s_i), \Psi_{\frac{m}{3\alpha}}(s_i), \Psi_{\frac{m}{4\alpha}}(s_i), \Psi_{\frac{m}{5\alpha}}(s_i)]$. Nos experimentos foi utilizado $\alpha = 5$.

Para ilustrar, a Fig. 5.5 mostra os valores de curvatura dos pontos salientes da forma da Fig. 5.5a. Para os pontos de saliências côncavas (1,4,5,8) a curvatura é positiva, enquanto os pontos de saliências convexas (2,3,6,7) tem curvaturas negativas. Quando o valor de t aumenta, obtém-se uma representação mais global da forma. Por outro lado, quando t diminui obtemos uma representação mais local. Por exemplo, considerando a saliência 7, ela é um ponto de saliência convexa para pequenos valores de t , no entanto, torna-se côncava quanto t cresce, porque a região considerada no cálculo também se expande.

Figura 5.5: (a) Uma forma com suas saliências; (b) Representação gráfica dos valores de curvatura das saliências usando diferentes valores de t na equação da curvatura (Eq. 5.1). O valor de m representa a quantidade de pontos do contorno.

5.3.3 Geração do Dicionário de Curvaturas

Antes de atribuir uma palavra-visual para cada ponto saliente, primeiro é necessário obter um *Dicionário de Curvaturas*, que é peça-chave da modelagem *Bag-of-Visual-Words*. O descritor BoSP proposto também se utiliza da mesma estratégia utilizada na metodologia *Bag-of-Visual Words* para gerar o dicionário: agrupamento das características extraídas de um sub-conjunto de imagens da base de dados.

Formalmente, seja $P = \{s_1, s_2, \dots, s_q\}$ o conjunto das saliências extraídas de um sub-conjunto da base de dados, em que cada saliência s_i é representada pelo vetor-de-características apresentado na subseção anteriormente. O Dicionário de Curvaturas é dado pela divisão de P em k distintos *clusters* $\pi_k = \{C_1, C_2, \dots, C_k\}$, tal que $C_1 \cup C_2 \cup \dots \cup C_k = P$, $C_i \neq \emptyset$, e $C_i \cap C_j = \emptyset$ para $i \neq j$ e $i, j = 1 \dots k$. Uma palavra-visual w_i é o centróide do *cluster* C_i .

O tamanho do dicionário é o número k de *clusters* fornecidos ao algoritmo de agrupamento, em que k é um número determinado empiricamente. No entanto, como será visto na seção experimental, um tamanho do dicionário muito grande tem pouco efeito

na qualidade da descrição e na recuperação de formas, mas afeta a dimensionalidade do vetor-de-características final e a eficiência do sistema de recuperação.

5.3.4 Descrição das Saliências em Palavras-Visuais

Uma vez que o Dicionário de Curvaturas foi gerado, o próximo passo consiste em atribuir uma palavra-visual para cada ponto saliente extraído e representado usando os passos descritos anteriormente. Neste trabalho, a palavra-visual de um ponto saliente s_i é a palavra-visual $(w_j)_{j=1\dots k}$ mais próxima. Dois pontos salientes da mesma forma podem ser atribuídos à mesma palavra, e uma palavra do dicionário pode não ser atribuída a nenhum ponto.

A principal ideia de usar um dicionário é considerar que os pontos de saliências são semelhantes às palavras-textuais presentes em um documento textual. Portanto, uma forma será composta por palavras-visuais como um documento textual é composto por palavras-textuais. Esta analogia torna fácil e simples a tarefa de calcular a similaridade.

Obtendo o Histograma Global

O Histograma Global é obtido pela simples contagem (quantidade) em que cada palavra-visual ocorre na forma. Isso significa que, para cada palavra-visual $(w_j)_{j=1\dots k}$, do Dicionário de Curvaturas, conta-se quantas vezes esta palavra aparece na forma, gerando um histograma de palavras-visuais. Se apenas o Histograma Global for usado para representar a forma, a dimensionalidade final do vetor-de-característica seria k , em que k é o número de palavras-visuais do dicionário.

A principal limitação da utilização de apenas o Histograma Global como a descrição final da forma é que apenas as estatísticas de primeira ordem seriam codificadas, ignorando assim o fato de que as estatísticas espaciais, tais como a distribuição de distâncias, são importantes para o reconhecimento de formas. Por esta razão, foi proposta uma nova maneira de codificar o arranjo espacial das palavras-visuais e, assim, enriquecer a descrição final da forma e desse modo melhorar a discriminação das imagens.

Obtendo os Histogramas Espaciais

Em vez de calcular somente um histograma, a proposta é calcular Z histogramas (para $Z > 1$) para codificar o arranjo espacial das palavras-visuais. Cada histograma representará a distribuição das distâncias visuais das palavras em relação ao centróide da forma. O centróide é o ponto de centro geométrico da forma, invariante a transformações geométricas, e que pode ser facilmente obtido pelos valores médios da coordenadas x e y dos pontos do contorno.

Para calcular a distribuição de distância, primeiro é necessário determinar a distância d do ponto saliente mais distante a partir do centróide, para determinar a zona de fronteira.

Então, define-se Z zonas equidistantes de distribuição, e então conta-se as ocorrências das palavras-visuais em cada zona.

A Figura 5.6 ilustra a ideia proposta utilizando $Z = 3$ zonas de distribuições de distância e considerando a distância Euclidiana. Esta estratégia é muito útil para distinguir formas com a mesma quantidade de palavras-visuais, uma vez que duas formas podem ter o mesmo Histograma Global, mas não a mesma distribuição de distâncias de palavras-visuais.

Figura 5.6: Para codificar o arranjo espacial das palavras visuais, o espaço da forma é dividido em zonas de acordo com a distância a partir do centróide e então calcula-se um histograma em cada zona. Neste exemplo, utilizou-se três zonas e distância Euclidiana.

O valor de Z é determinado empiricamente. No entanto, como será mostrado na seção experimental, para valores grandes de Z a acurácia no reconhecimento de formas não aumenta.

Representação Final

O vetor-de-características que representa a descrição final da forma é uma concatenação do Histograma Global e dos Histogramas Espaciais obtidos na etapa anterior. A dimensionalidade deste vetor-de-característica é $k(1 + Z)$, em que k é a quantidade de palavras do Dicionário de Curvaturas e Z o número de zonas de distribuições de distância consideradas no cálculo dos Histogramas Espaciais.

No entanto, o Histograma Global pode ser obtido pela soma de todos os Histogramas Espaciais. Sendo assim, com esta estratégia a dimensionalidade final do vetor-de-características pode ser reduzido para kZ , que é mais compacto. Assim, agora que todos os vetores-de-características têm a mesma dimensionalidade, pode-se usar funções tradicionais para calcular a distância entre eles e medir a similaridade.

A distância tradicional para medir a similaridade entre duas formas A e B representados por seus histogramas de palavras-visuais $H^A = \{h_1^A, h_2^A, \dots, h_l^A\}$ e $H^B = \{h_1^B, h_2^B, \dots, h_l^B\}$ é a distância L_1 ou City-Block:

$$D_{L_1}(H^A, H^B) = \sum_{i=1}^l |h_i^A - h_i^B|$$

Note que, a distância L_1 é utilizada pois é computacionalmente mais rápida que a distância Euclidiana.

5.4 Usando Características Globais

Com o objetivo de aumentar o poder discriminativo no reconhecimento e recuperação de formas, neste trabalho é proposta a estratégia de utilizar características globais no cálculo da similaridade entre duas formas. Uma característica global analisa a imagem como um todo, e esta é uma informação preciosa para aumentar a distância entre formas muito diferentes e reduzir a similaridade entre formas visualmente muito semelhantes.

Para medir os aspectos globais, neste trabalho foram consideradas duas características geométricas importantes: excentricidade e solidez:

- A excentricidade E é uma medida de relação de aspecto. Ela é calculada pelos principais eixos do mínimo retângulo delimitador na forma. Este valor é a razão entre o comprimento do eixo menor com o eixo maior e o valor pode variar entre 0 e 1. Um quadrado, por exemplo, tem excentricidade igual a 1, enquanto que um segmento de reta tem excentricidade igual a 0.
- A solidez S especifica a percentagem dos pixels da região convexa (Convex Hull) da forma. *Convex Hull* é uma matriz que especifica o menor polígono convexo que pode conter a forma. A solidez é calculada como a relação da área da forma com sua região convexa.

Para ilustrar, a Figura 5.7 mostra o mínimo retângulo delimitador e o *Convex Hull* de uma dada forma.

Figura 5.7: Mínimo retângulo delimitador e *Convex Hull* de uma forma.

Sejam H^A e H^B , respectivamente, os histogramas das formas A e B , a distância final para calcular a similaridade entre essas duas formas considerando suas características globais (solidez S e excentricidade E), é definida por:

$$D(A, B) = D_{L_1}(H^A, H^B) + \beta \quad (5.3)$$

em que:

$$\beta = |S_A - S_B| + |E_A - E_B| \quad (5.4)$$

Duas formas A e B são consideradas semelhantes quando $D(H^A, H^B) \rightarrow 0$. Se as formas A e B possuem valores de excentricidade e solidez semelhantes, então $\beta \rightarrow 0$ e uma pequena penalidade na distância final de similaridade. Esta estratégia contribui para aumentar a capacidade de distinguir formas visualmente diferentes.

A Figura 5.8 mostra um exemplo em que estas duas características são úteis ao se calcular a similaridade entre as formas: nota-se que a forma A é visualmente mais semelhante à forma B do que a C, no entanto B tem mais saliências do que a forma C. Esta situação pode penalizar negativamente a distância final ao se analisar apenas os pontos de saliências entre as formas. Para contornar tal problema, as características globais contribuem para diminuir a similaridade nessa situação, uma vez que os valores de excentricidade e solidez das forma A e B são menores do que as formas A e C.

Figura 5.8: A forma A é visualmente mais semelhante a forma B do que a C. No entanto, a forma B tem mais pontos salientes do que C, o que pode penalizar negativamente a similaridade. Duas características globais (excentricidade e solidez) são adicionadas para reduzir essa diferença.

5.5 Avaliações Experimentais

Essa seção apresenta os testes experimentais realizados comparando o descritor proposto BoSP com outros da literatura, usando diferentes bases de imagens. A versão do descritor proposto que considera as características globais na análise de similaridade será referenciado como BoSP+GF (*Bag-of-Saliency-Points + Global Features*).

5.5.1 Bases de Imagens

Para os experimentos, foram utilizadas três diferentes bases clássicas da literatura:

- MPEG-7 CE-Shape-1 (Bober, 2001): este é um dos conjuntos de dados mais utilizado e avaliado para a recuperação de formas. Ele é composto por 1.400 formas

de diferentes tipos, classificados em 70 classes com 20 exemplos em cada classe. O reconhecimento de formas usando esse banco de dados não é uma tarefa simples, uma vez que existem formas que são visualmente diferentes de outros membros de sua própria classe. Além disso, há formas que são altamente semelhantes a exemplos de outras classes;

- Kimia-216: é composto por 18 classes cada uma das quais abrangendo 12 formas, em que algumas delas são parcialmente oclusas. Essa base contém silhuetas de aves, ossos, tijolos, camelos, carros, crianças, postais clássicos, elefantes, entre outros.
- Kimia-99: tem 9 classes cada uma composta por 11 formas. Essa base contém silhuetas de coelhos, quadrúpedes, homens, aviões, peixes, mãos, raios, ferramentas e uma classe de diversas formas variáveis.

5.5.2 Impacto dos Parâmetros

O descritor proposto necessita de dois parâmetros a serem definidos: o tamanho do Dicionário de Curvaturas k e o número de zonas de distribuições de distância Z . Para investigar os melhores valores destes dois parâmetros, foram realizados exaustivamente vários experimentos com valores diferentes, e as Figuras 5.9-a-b-c mostram os valores da Precisão média (mAP - *mean Average Precision*) obtidos nestes experimentos para as bases avaliadas.

Analizando os resultados, um dicionário com tamanho superior a 20 palavras não obtém uma melhoria global na precisão em nenhuma das bases avaliadas. Nota-se também que a quantidade de zonas interfere na acurácia do reconhecimento, no entanto, uma quantidade de zonas maior do que 4 não melhora a precisão dos resultados. Portanto, com base nesses experimentos, os valores apropriados de k e Z foram definidos como 15 e 4, respectivamente, em todas as bases.

5.5.3 Resultados Obtidos

Recuperação de Formas

A medida de avaliação adotada na literatura para avaliar o desempenho de descritores na base MPEG-7 é conhecida como “*Bull’s eye score*” ou Revocação-40. Essa medida consiste em contar quantas imagens, dentre as 40 primeiras imagens do *ranking*, pertencem à classe da imagem de consulta. Como a base MPEG-7 tem 1400 imagens divididas igualmente em 70 classes, se forem usadas todas as imagens da base como imagem de consulta, então o total de acertos possíveis é 28000 (1400×20). A medida *Bull’s eye* é a razão dos acertos obtidos pelo valor máximo de acertos possíveis.

Figura 5.9: mAP (*mean Average Precision*) valores obtidos usando tamanhos de dicionários diferentes e diferentes quantidades de zonas de distribuição de distância. (a) MPEG-7 (b) Kimia-216 (c) Kimia-99. Para $Z = 1$ considera-se apenas o Histograma Global.

A Tabela 5.1 mostra os valores da medida *Bull's eye score* reportados por vários trabalhos na literatura, assim como o tempo médio utilizado por eles para comparar duas formas. Embora a taxa de recuperação do descritor proposto não ser tão boa como as dos descritores propostos por (Bai et al., 2014; C. Xu e Tang, 2009; J. Wang e Latecki, 2012; M. Liu e Nielsen, 2010, 2012; Nguyen e Porikli, 2013; R. Gopalan e Chellappa, 2010; Wang e Liang, 2010), o descritor proposto supera esses descritores em termos de eficiência de tempo na análise da similaridade, pois o descritor proposto utiliza uma métrica simples para calcular a distância entre dois vetores (distância L_1). Os descritores que tem taxas de recuperação superiores ao descritor proposto utilizam técnicas de similaridade mais sofisticadas, tais como o método de Hungarian (S. Belongie e Puzicha, 2002), programação dinâmica (Nguyen e Porikli, 2013) e divergência de Bregman (M. Liu e Nielsen, 2010), que muitas vezes gastam um tempo computacional considerável para analisar a similaridade entre duas formas.

Tabela 5.1: Taxa *Bulls-Eye Score* e o tempo médio gasto na análise da similaridade para a base MPEG-7.

Descritor	Taxa (%)	Tempo (ms)
Fourier	30,9	0,035
Moment Invariant (MI)	38,6	0,01
MS-Fractal (Torres et al., 2003)	54,8	0,029
TSDIZ (Andaló et al., 2010)	69,38	N/A
Zernike Moments (Khotanzad e Hong, 1990)	70,22	N/A
Curvature Scale Space (Mokhtarian e Suomela, 1998)	75,44	N/A
Visual Parts (Latecki e Lakamper, 2000)	76,45	N/A
Shape Context (S. Belongie e Puzicha, 2002)	76,51	N/A
SSD (Pedrosa et al., 2013)	79,77	0,37
★ BoSP	80,9	0,021
Biswas (no rank refinement) (S. Biswas e Chellappa, 2010)	81,8	N/A
BAS (Arica e Vural, 2003)	82,37	N/A
Planar Graph Cuts (F. R. Schmidt e Cremers, 2009)	85	N/A
Inner Distance Shape Context (DP) (Ling e Jacobs, 2007)	85,4	N/A
Symbolic Representation (Daliri e Torre, 2008)	85,92	N/A
★ BoSP+GF	86,35	0,022
ROMS (DP) (Z. Li e Qi, 2012)	86,36	N/A
Biswas (with rank refinement) (S. Biswas e Chellappa, 2010)	86,48	N/A
Triangle Area Representation (N. Alajlan e Freeman, 2008)	87,13	N/A
Shape Tree (Felzenszwalb e Schwartz, 2007)	87,7	N/A
GMM (tBD hard clustering) (M. Liu e Nielsen, 2010)	89,1	N/A
Contour Flexibility (C. Xu e Tang, 2009)	89,31	N/A
Locally Affine Invariant (Wang e Liang, 2010)	89,62	N/A
Height Functions (DP) (J. Wang e Latecki, 2012)	89,66	N/A
Height Functions (DP + DCN) (J. Wang e Latecki, 2012)	90,35	N/A
Shape Vocabulary (Bai et al., 2014)	90,41	N/A
SVS (DP) (Nguyen e Porikli, 2013)	91,07	N/A
GMM (tBD soft clustering) (M. Liu e Nielsen, 2012)	93,41	N/A
AIR (R. Gopalan e Chellappa, 2010)	93,67	N/A

Para as bases Kimia-99 e Kimia-216, o descritor proposto foi comparado com quatro outros descritores representativos da literatura: MI (Momentos Invariantes), Fourier (Descritores de Fourier), MS Fractal (Multi-scale Fractal Dimension) (Torres et al., 2003) e SSD (Shape Saliency Descriptor) (Pedrosa et al., 2013). A Figura 5.11-ab mostra as curvas de PR (Precisão vs. Revocação) para as bases Kimia-99 e Kimia-216. O descritor BoSP+GF apresentou a melhor curva em comparação com os descritores avaliados. O descritor BoSP proposto obteve um desempenho semelhante ao descritor SSD, porém sendo 53% mais rápido ao calcular a similaridade de formas. A Figura 5.10 apresenta um exemplo do resultado da recuperação usando os descritores avaliados: as imagens dentro da caixa não pertence à classe da imagem consulta. Nota-se que o descritor proposto apresenta um resultado visual melhor do que os outros descritores avaliados, relatando uma alta qualidade de reconhecimento de formas.

Figura 5.10: Uma amostra do *ranking* gerado pela recuperação usando os descritores avaliados.

Avaliação Qualitativa

Um aspecto importante a ser considerado nessas avaliações é o fato de que o reconhecimento de formas usando a base MPEG-7 não é trivial, uma vez que algumas formas são visualmente diferentes de outros membros de sua própria classe. Por outro lado, existem formas altamente semelhantes, num sentido visual, com exemplos de outras classes. Devido a este fato, foi realizada uma avaliação qualitativa para analisar se o descritor proposto tem uma precisão subjetiva maior do que a sua precisão quantitativa. O objetivo é verificar se o descritor proposto traz satisfação subjetiva para o usuário do sistema de recuperação. Para isso, foi realizado um experimento qualitativo com 23

Figura 5.11: Curvas de PR para as bases: (a) Kimia-216 (b) Kimia-99.

peças, que analisaram o *top* 20 de 150 diferentes imagens de consulta. Foi pedido que os usuários selecionassem as imagens não relevantes em relação à imagem-consulta.

O gráfico da Figura 5.12 mostra os valores de precisão obtidos até o *top*-20 no *ranking*, considerando: a avaliação humana (qualitativa) e a classificação do conjunto de dados (quantitativo) já definida com a base. Com relação a avaliação qualitativa, o descritor BoSP+GF apresentou uma precisão subjetiva com precisão 8% maior do que a sua precisão quantitativa. Este resultado demonstra que o descritor proposto é capaz de capturar características visuais semelhantes a formas de outras classes da imagem-consulta, mesmo que as formas possuam um número diferente de pontos salientes.

Figura 5.12: Precisão do descritor BoSP+GF para a avaliação humana (qualitativo) e para a classificação definida da base (quantitativo).

5.6 Considerações Finais

Neste capítulo foi apresentado um novo descritor, chamado de *Bag-of-Saliency-Points* (BoSP), para a recuperação de imagens baseada em formas usando pontos salientes como

características de descrição. Pontos salientes são muito úteis na descrição da imagem, porque eles codificam as partes visuais mais importantes da imagem de uma maneira compacta e invariante a transformações geométricas, tais como rotação e translação.

O fato é que duas formas diferentes podem ter quantidades diferentes de pontos salientes e, por isso, é desafiadora a tarefa de medir a similaridade entre formas analisando apenas suas saliências. Porém, a abordagem proposta utiliza a estratégia baseada em *Bag-of-Visual-Words* para representar os pontos salientes em um único vetor-de-características. Esta estratégia permite calcular a similaridade de formas usando funções de distância rápidas, tais como as distâncias Euclidiana e/ou City-block.

A ideia adotada pelo descritor BoSP é considerar cada tipo de saliência como uma palavra-visual. Para isso, cada saliência é primeiramente representada através de seus valores de curvatura usando uma abordagem em multi-escala e, posteriormente, atribui-se uma palavra-visual para cada saliência de acordo com um Dicionário de Curvaturas, ambas abordagens propostas neste trabalho. Em seguida, conta-se o número de palavras-visuais que ocorrem na forma obtendo um Histograma Global. Para enriquecer a representação final, também foram propostos os Histograms Espaciais que codificam a distribuição das distâncias visuais das palavras em relação ao centróide da forma. Estes Histogramas Espaciais produzem uma descrição mais robusta e discriminatória, uma vez que duas formas podem ter o mesmo Histograma Global, mas não a mesma distribuição das palavras-visuais.

Também foi apresentada uma versão estendida do descritor BoSP com características globais, referenciado como BoSP+GF. Esta versão estendida utiliza duas características globais para enriquecer a descrição semântica das formas, aprimorando de uma maneira eficiente a precisão na recuperação sem pagar computacionalmente mais por isso. Essa versão requer apenas dois valores adicionais no vetor-de-características final, o que não afeta muito a dimensionalidade nem o espaço de armazenamento requerido.

A técnica proposta foi comparada com vários outros descritores da literatura e avaliados utilizando três bases de dados representativas da área. As avaliações experimentais indicaram que o principal ganho do descritor proposto é a alta qualidade no reconhecimento visual e uma rápida medida de similaridade, mostrando que o descritor BoSP é uma contribuição valiosa para a recuperação de imagens baseada em formas.

Consulta por Similaridade Apoiada por Imagens Dominantes

6.1 Considerações Iniciais

A subjetividade do conceito de similaridade é uma tarefa muito complexa de se modelar computacionalmente, uma vez que a expectativa de diferentes usuários pode ser distinta, ou mesmo a necessidade do próprio usuário pode ser diferente em momentos distintos. Quando o usuário submete uma consulta em um sistema CBIR, a expectativa é que este sistema retorne apenas imagens de seu interesse (imagens relevantes) dentro de um *contexto semântico* esperado por ele, o que nem sempre é possível capturar somente usando as características intrínsecas extraídas das imagens. Isso causa uma separação entre o que um indivíduo reconhece em uma imagem, e o que pode ser esperado de um sistema computacional que tenta reproduzir esse reconhecimento, originando o problema do *gap* semântico.

Para reduzir o problema do *gap* semântico e aproximar a recuperação à expectativa do usuário, na prática os sistemas de recuperação se apoiam em algum método de *pós-processamento* aplicado na caracterização da imagem e/ou no momento da consulta. Por exemplo, técnicas de seleção de características e ponderação de atributos são soluções bastante utilizadas para enriquecer a caracterização final da imagem e a análise da similaridade. É muito comum também utilizar interações com o usuário (*relevance feedback*) no momento da consulta, capturando a necessidade do usuário em tempo real (*on the fly*).

O fato é que, em geral, os sistemas de busca por conteúdo realizam a análise de imagens em pares, isto é, o cálculo da similaridade é realizado considerando apenas duas

imagens: a imagem-consulta e uma imagem da base de dados. O *ranking* final é obtido considerando apenas o cálculo dessa similaridade, ignorando informações valiosas que podem estar codificadas nas relações entre as várias imagens da base de dados. Assim, uma maneira de se aproximar um sistema CBIR à expectativa do usuário é analisar a similaridade entre as imagens baseada na *informação contextual* da aplicação.

Com o objetivo de diminuir o *gap* semântico, esse capítulo apresenta uma nova abordagem de consulta por similaridade, denominada de *Similaridade Baseada em Imagens Dominantes* ($SimD_{Im}$), visando ampliar a precisão da recuperação analisando o contexto da coleção de imagens. Essa nova abordagem se apoia em um novo conceito de similaridade chamado de *Imagens Dominantes*, que é um sub-conjunto reduzido da coleção de imagens formado pela exclusão daquelas imagens consideradas muito similares a partir de uma imagem de referência. Esse novo conceito permite eficácia e eficiência quando se deseja analisar a informação contextual da aplicação, que é a proposta da abordagem $SimD_{Im}$.

Em suma, a técnica que será apresentada consiste em dois módulos: *mining* (mineração) e *matching* (casamento). O módulo de *mining* consiste em extrair as *Imagens Dominantes* a partir do *ranking* inicial gerado por uma imagem-consulta. Já o módulo de *matching* consiste em re-fazer a busca por conteúdo utilizando a informação contextual obtida com o conjunto de *Imagens Dominantes* na fase de mineração. A tarefa de *matching* pode ser realizada de duas maneiras: não-supervisionada ou supervisionada com a interação do usuário. A nova proposta que será apresentada permite enriquecer a análise da similaridade em consultas por conteúdo, aprimorando a precisão da recuperação com resultados mais próximos ao esperado pelo usuário.

6.2 Contextualização

Uma das estratégias mais comuns na literatura para reduzir o problema do *gap* semântico consiste no aprendizado de características que são mais representativas para as consultas. A ideia é aprimorar as consultas por similaridade fazendo com que, de alguma maneira, as distâncias entre imagens semelhantes sejam menores e as distâncias entre imagens não semelhantes sejam maiores. As principais abordagens se dividem em técnicas que atuam sobre as *funções de similaridade* e técnicas que atuam sobre o *espaço de dados*. As técnicas que consideram funções de similaridade visam a calibrar essas funções por meio de vetores-de-pesos, e as técnicas de transformação do espaço consideram metodologias de seleção e transformação de características produzindo um conjunto completamente novo de características para representar os dados.

Na área textual, por exemplo, é muito comum o uso de *métodos de ponderação* na mineração de textos com o objetivo de determinar os melhores termos para a modelagem de documentos (Yang e Pedersen, 1997). Esses métodos podem ser divididos em 3

categorias, de acordo com o tipo de medida de relação entre o documento e atributo no qual o método se baseia: frequência de termos, variância de termos e contexto. A frequência de termo, conhecida como TF (*Term Frequency*), e a frequência de documentos, conhecida como DF (*Document Frequency*), são duas medidas de relação entre atributos e documentos facilmente calculadas com baixo custo computacional. A TF contabiliza a frequência absoluta de um determinado termo ao longo da coleção de documentos e a DF contabiliza o número de documentos que um determinado termo aparece. A TF-IDF (*Term Frequency-Inverse Document Frequency*) é uma das medidas mais tradicionais e utilizadas. A ideia é ponderar a TF dos termos em função de suas distribuições ao longo da coleção, dando menor peso àqueles termos que aparecem em muitos documentos. Para isso é introduzido o valor do inverso da DF, obtendo a IDF (*Inverse Document Frequency*) calculado por: $IDF_j = \log(N/DF_j)$, na qual DF_j é a frequência de documento do j -ésimo atributo e N a quantidade de documentos da coleção. A TF-IDF é obtida pela multiplicação da TF pela IDF.

Por outro lado, existem abordagens que realizam a *seleção de características*, ou seja, recebem um vetor-de-características como entrada e o objetivo é obter um subconjunto de características desse vetor que aprimore a precisão do sistema pela eliminação de características pouco preditivas. O subconjunto gerado pode ser: um subconjunto de características com um tamanho específico que otimiza o desempenho do sistema ou um subconjunto com o melhor equilíbrio entre o desempenho e acurácia na recuperação/classificação. Essas técnicas podem ser divididas em três categorias: filtros, *wrappers* e métodos híbridos (Yun et al., 2007). Métodos de filtragem são aplicados diretamente aos conjuntos de dados e geralmente fornecem um índice para a qualidade de cada característica usando alguma medida baseada na propriedade dos dados. *Wrappers* e métodos híbridos usam a precisão de classificação (ou outra medida de desempenho) de algum algoritmo de aprendizagem para orientar um processo de pesquisa no espaço de características e utilizar os parâmetros desses algoritmos de aprendizagem para avaliar a qualidade de cada característica ou conjunto de características.

Além da seleção de características, uma maneira mais efetiva de analisar o conteúdo visual de uma imagem é usar uma *combinação de características* (Mansano et al., 2012). A ideia é que as informações de diferentes fontes possam quantificar melhor a percepção similar. Assim, diferentes descritores podem gerar diferentes *rankings*, gerando informações complementares. As abordagens que combinam dados de diferentes *rankings* são chamadas de *rank aggregation*. Essas técnicas consideram as posições das imagens em diferentes *rankings* gerados em múltiplas recuperações, a fim de obter um *ranking* único com desempenho mais acurado (Shaw et al., 1994).

Porém, a utilização por si só das características extraídas de uma imagem, não garante um bom resultado na análise da similaridade do conteúdo visual entre imagens. Em alguns casos, o desempenho depende também da função da distância utilizada para medir essa

similaridade. O trabalho de (Papa et al., 2009), por exemplo, mostra que se pode melhorar até 7,28% a precisão do reconhecimento em tarefas de classificação de imagem, para os mesmos conjuntos de treinamento e teste, apenas mudando a função distância.

Técnicas de *Realimentação de Relevância* também tem sido exploradas para atuar no processo de recuperação. Essas técnicas possibilitam ao usuário indicar quais instâncias do conjunto retornado por uma consulta são relevantes e/ou não-relevantes. O trabalho de (Rocchio, 1971) descreve uma das primeiras técnicas de realimentação de relevância, chamada Alteração do Centro de Consulta (*Query Point Movement*). Essa técnica consiste em mudar o elemento central da consulta, escolhendo um novo elemento que seja mais próximo dos objetos relevantes e mais distante dos objetos não-relevantes. Formalmente, dado o vetor-de-características $\vec{v}_0$ utilizado na consulta, o novo vetor $\vec{v}_1$ é obtido pela equação:

$$\vec{v}_1 = \vec{v}_0 + \frac{1}{N_R} \sum_{i=1}^{N_R} \vec{r}_i - \frac{1}{N_I} \sum_{i=1}^{N_I} \vec{s}_i \quad (6.1)$$

em que, N_R e N_I referem-se, respectivamente, à quantidade de imagens relevantes e não-relevantes definidas pelo usuário, $\vec{r}_i$ e $\vec{s}_i$ são os vetores das imagens relevantes e não-relevantes respectivamente. O problema dessa abordagem é não considerar a proporção entre as distâncias, pois imagens muito distantes podem ocasionar um deslocamento muito grande no vetor de consulta, prejudicando o processo de recuperação.

Geralmente, a consulta pelas imagens mais próximas à uma dada imagem-consulta leva em consideração apenas a relevância individual de cada imagem computada pela função de distância em relação à imagem-consulta. Assim, essa consulta não é capaz de considerar o relacionamento dos elementos entre si, o que possibilitaria aumentar a informação na análise da similaridade entre imagens, que é a ideia da proposta que será apresentada a seguir.

6.3 Método Proposto: Similaridade Baseada em Imagens Dominantes

O método proposto, denominado de *Similarity Based on Dominant Images* ($SimD_{Im}$), consiste em enriquecer semanticamente a consulta por similaridade entre imagens analisando o contexto da aplicação, ou seja, a *coleção de imagens* utilizada na recuperação. Para isso, neste trabalho é introduzido um novo conceito de similaridade denominado de *Imagens Dominantes*, que é um sub-conjunto menor e que representa, de uma maneira mais diversificada, toda a coleção de imagens recuperadas na consulta.

A seguir é apresentada uma definição formal do problema, seguido do conceito de *Imagens Dominadas* e *Imagens Dominantes*, e as etapas do método $SimD_{Im}$ para: (i) a

extração das *Imagens Dominantes*; e (ii) a função de distância para medir a similaridade contextual entre duas imagens da coleção.

6.3.1 Definição Formal do Problema

Seja $\mathcal{C} = \{img_1, img_2, \dots, img_N\}$ uma coleção de imagens, e $\mathcal{D}$ um descritor definido por uma tupla (ϵ, δ_{sim}) em que:

- $\epsilon : \mathcal{C} \rightarrow \mathbb{R}^n$ é uma função que extrai um vetor-de-características $\vec{v}$ de uma imagem $img \in \mathcal{C}$;
- $\delta_{sim} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ é uma função que computa a distância (similaridade) entre duas imagens descritas por seus vetores-de-características.

A fim de obter a similaridade entre duas imagens img_i e img_j é necessário computar o valor $\delta_{sim}(\epsilon(img_i), \epsilon(img_j))$, porém, por simplicidade será usada a notação $\delta_{sim}(img_i, img_j)$.

Dada uma imagem-consulta img_q , um *ranking* $R_q = \{img_1, img_2, \dots, img_N\}$ é obtido computando-se a distância de img_q para todas as imagens da coleção, de tal modo que $\delta_{sim}(img_q, img_1) \leq \delta_{sim}(img_q, img_2) \leq \dots \leq \delta_{sim}(img_q, img_N)$. O objetivo desse trabalho é obter um novo *ranking* $\tilde{R}_q$ que seja semanticamente superior a R_q , ou seja, um novo *ranking* $\tilde{R}_q$ cuja ordem das imagens atenda melhor às expectativas do usuário. Para gerar $\tilde{R}_q$, neste trabalho é proposta uma nova função de similaridade $\delta_{simD_{Im}}$ baseada nos conceitos de *Imagens Dominadas* e *Imagens Dominantes* (D_{Im}) que são apresentados a seguir.

6.3.2 Definição de Imagens Dominadas e Dominantes

Pode-se considerar que duas imagens img_i e img_j são similares se $\delta_{sim}(img_i, img_j) < \xi$, tal que ξ é uma constante que indica uma distância máxima de similaridade, ou seja, um valor máximo permitido para que duas imagens sejam consideradas similares. Desse modo, dado um valor de ξ , a coleção de imagens $\mathcal{C}$ pode ser reduzida a um sub-conjunto menor $\mathcal{C}' \subset \mathcal{C}$ de imagens similares, ou seja, $|\mathcal{C}'| \ll |\mathcal{C}|$ tal que, dada uma imagem $img_i \in \mathcal{C}$ existe uma imagem $img_j \in \mathcal{C}'$ em que $\delta_{sim}(img_i, img_j) < \xi$.

A Figura 6.1 ilustra um exemplo intuitivo da ideia. A Figura 6.1-a mostra uma coleção de imagens $\mathcal{C}$ em um espaço de características bi-dimensional. A Figura 6.1-b mostra a ligação entre imagens que são similares entre si dado um valor ξ de similaridade. Por exemplo, a imagem img_1 é similar às imagens img_2 e img_4 . A Figura 6.1-c mostra uma sub-coleção de imagens menor formada por imagens que não são similares entre si. Neste exemplo, o conjunto inicial de 7 imagens foi reduzido à sub-conjunto de apenas 3 imagens.

Contudo, para se obter $\mathcal{C}'$ é necessário definir uma distância de similaridade ξ . Para obter $\mathcal{C}'$, independente de parâmetro, pode-se assumir que ξ é um valor relativo, ou seja,

Figura 6.1: Exemplo mostrando um subconjunto formado por imagens não-semelhantes, dada uma distância máxima de similaridade.

que ele depende de uma imagem de referência. Assim, para encontrar a distância ξ pode-se usar a relação de distância entre as imagens para uma imagem de referência, analisando as imagens ao redor. Essa imagem de referência pode ser qualquer imagem da coleção, por exemplo, a imagem-consulta. Desse modo, não é necessário especificar um parâmetro ξ que indique o grau de similaridade entre duas imagens, basta dizer que uma imagem é mais similar a outra em relação à imagem-consulta. Esse é o princípio das Definições de *Imagem Dominada* e *Imagens Dominantes*

Definição 6.3.1 Dada uma imagem-consulta img_q e uma coleção de imagens $\mathcal{C}$, define-se **Imagem Dominada** como sendo uma imagem $img_i \in \mathcal{C}$, tal que existe uma imagem $img_j \in \mathcal{C}$ que satisfaz as seguintes condições de similaridade:

1. $\delta_{Sim}(img_i, img_j) < \delta_{Sim}(img_i, img_q)$
2. $\delta_{Sim}(img_i, img_j) < \delta_{Sim}(img_j, img_q)$

Em outras palavras, a Definição 6.3.1 diz que uma imagem $img_i \in \mathcal{C}$ é dita *dominada*, se existir uma outra imagem $img_j \in \mathcal{C}$ em que: (i) img_i é mais similar à img_j do que img_i à imagem-consulta e (ii) img_i é mais similar à img_j do que img_j à imagem-consulta. A segunda restrição impede que exista apenas uma única imagem img_j dominando todas as outras imagens conjunto, pois como a relação de dominância é dada pela distância relativa das imagens à imagem-consulta img_q , sem a segunda restrição seria possível que uma imagem muito próxima de img_q domine todo o conjunto, assim img_j seria um *elemento único* que dominaria todas as outras imagens do conjunto, o que não permitiria ter uma visão abrangente da coleção de imagens. A Figura 6.2 ilustra um exemplo de imagem dominada.

A partir da definição de Imagens Dominadas, pode-se criar um conjunto formado por imagens que *não são dominadas* por nenhuma outra imagem do próprio conjunto. Essa é a ideia do conceito de Imagens Dominantes, dado a seguir.

Figura 6.2: A imagem i é dita imagem dominada, pois a distância de i para j é menor que i para a imagem q , além disso a distância de j para q é maior que a distância de j para i .

Definição 6.3.2 Define-se **Imagens Dominantes** como sendo um sub-conjunto $D_{Im} \subset \mathcal{C}$, em que não existe nenhuma imagem $img_i \in D_{Im}$ que seja dominada por outra imagem $img_j \in D_{Im}$.

A principal vantagem da construção do conjunto de Imagens Dominantes é reduzir uma grande coleção de imagens à um sub-conjunto menor e variado de imagens, e assim dar mais eficiência quando se deseja analisar o contexto da coleção, que é o objetivo desse trabalho.

6.3.3 Etapa de Mineração

Para gerar o conjunto D_{Im} de Imagens Dominantes a partir do conceito de Imagens Dominadas, neste trabalho é proposta a seguinte estratégia: dada uma imagem consulta img_q e seu *ranking* $R_q = \{img_1, img_2, \dots, img_N\}$, assume-se que $img_1 \in D_{Im}$, ou seja, o vizinho mais próximo. A intuição para considerar $img_1 \in D_{Im}$ no primeiro passo é assegurar que técnica inicie sua restrição de cardinalidade no contexto da imagem-consulta, fornecendo perspectivas da vizinhança e, posteriormente, de toda a coleção de imagens. Desse modo, dado que $\xi_1 = \delta_{Sim}(img_1, img_q)$, então pode-se analisar quais as outras imagens $img_i \in R_q$, para $i = 2 \dots N$, são dominadas por img_1 , verificando se: (i) $\delta_{Sim}(img_i, img_q) > \xi_1$ e (ii) $\delta_{Sim}(img_i, img_1) < \xi_1$. Se (i) e (ii) acontece, então a imagem é dominada por img_1 , logo $img_i \notin D_{Im}$, caso contrário $img_i \in D_{Im}$ e computa-se a distância $\xi_i = \delta_{Sim}(img_i, img_q)$ e continua a verificação para as outras imagens $img_j \in R_q$, para $j = i + 1 \dots N$. Tal procedimento é descrito em detalhes pelo Algoritmo 1.

A Figura 6.3 ilustra os passos do Algoritmo 1 para extrair as Imagens Dominantes de um espaço de características hipotético em 2D dado pela Figura 6.3a e utilizando a distância Euclidiana. Dada uma imagem consulta img_q e seu respectivo *ranking* de imagens similares R_q , o primeiro passo é adicionar a imagem mais similar img_1 ao conjunto D_{Im} e, em seguida, procurar quais imagens do *ranking* são dominadas por img_1 , no caso imagens img_2 e img_4 . Na segunda iteração, img_3 é adicionada ao conjunto D_{Im} , pois img_2 é dominada por img_1 e repete o processo verificando quais imagens são dominadas por img_3 (imagem img_5). Na terceira iteração, img_6 é adicionada ao conjunto D_{Im} e, novamente,

Algoritmo 1 Geração do conjunto de Imagens Dominantes**Entrada:** Imagem-consulta img_q e seu *ranking* $R_q = \{img_1, img_2, \dots, img_N\}$ **Saída:** Conjunto de Imagens Dominantes D_{Im}

```

1:  $D_{Im} = \emptyset$ 
2:  $D_{Aux} = \emptyset$  // conjunto auxiliar
3: para  $i = 1 : N$  faça
4:   se  $img_i \notin D_{Aux}$  então
5:      $img_i \rightarrow D_{Im}$  // adiciona o elemento ao conjunto
6:     para  $j = i + 1 : N$  faça
7:       se  $\delta_{Sim}(img_j, img_i) < \delta_{Sim}(img_j, img_q)$  e  $\delta_{Sim}(img_j, img_i) < \delta_{Sim}(img_i, img_q)$ 
8:         então
9:            $img_j \rightarrow D_{Aux}$  //  $img_i$  domina  $img_j$ , logo  $img_j$  nunca poderá ser dominante
10:        fim se
11:      fim para
12:    fim se
13:  fim para

```

verifica-se quais imagens são dominadas por img_6 (imagem img_7). Fim das iterações, as Imagens Dominantes neste exemplo são as imagens img_1 , img_3 e img_6 .

6.3.4 Etapa de Casamento

Após a etapa de mineração, toda a coleção de imagens da base de dados está agora representada por um conjunto *menor* e *diverso* de imagens que é o conjunto de Imagens Dominantes $D_{Im} = \{img_1, \dots, img_M\}$. A etapa de *matching* (casamento) consiste em considerar esse conjunto em uma nova análise de similaridade. A ideia é que o conjunto D_{Im} possui informações valiosas que auxiliam analisar o contexto da aplicação e, com isso, permite enriquecer a semântica envolvida na consulta por similaridade, gerando um novo *ranking* $\tilde{R}_q$ mais próximo da expectativa do usuário.

Dado que uma imagem dominante é uma imagem que “influencia” outras imagens, então a coleção de imagens pode ser dividida em dois grupos: *imagens dominantes* e *imagens dominadas*. A estratégia proposta para computar uma nova similaridade é considerar, para cada imagem dominada da coleção, sua respectiva imagem dominante. A ideia é que o grau de similaridade de uma imagem dominante se reflita para todas as suas imagens dominadas. Assim, se uma dada imagem dominante img_j é muito similar à imagem-consulta, então todas as imagens dominadas por img_j também podem ser. Essa é a estratégia adotada pela nova função de similaridade $\delta_{SimD_{Im}}$, dada pela Definição 6.3.3.

Definição 6.3.3 Dada uma imagem-consulta img_q e seu respectivo *ranking* $R_q = \{img_1, \dots, img_N\}$, um novo *ranking* $\tilde{R}_q$ é obtido considerando a distância de cada imagem $img_i \in R_q$ à sua respectiva Imagem Dominante $img_j \in D_{Im}$:

$$\delta_{SimD_{Im}}(img_i, img_q) = \delta_{Sim}(img_i, img_q) + \delta_{Sim}(img_j, img_q) \quad (6.2)$$

Figura 6.3: A figura (b) mostra os passos para a geração do conjunto de Imagens Dominantes do exemplo da figura (a). Dado um *ranking* inicial, a imagem mais próxima img_1 à imagem consulta img_q é inserida no conjunto (circulo amarelo), verifica-se quais imagens são dominadas por ela (círculos cinzas). Repete-se tal processo para outras imagens não dominadas e não inseridas no conjunto de Imagens Dominantes.

Pela Definição 6.3.3, as distâncias de cada imagem no *ranking* $R_q = \{img_1, \dots, img_N\}$ são atualizadas considerando a distância da sua imagem dominante img_j à imagem-consulta img_q , ou seja, $\delta_{Sim}(img_j, img_q)$. Assim, se $\delta_{Sim}(img_j, img_q) \rightarrow 0$, a nova distância é pouco penalizada. Porém, se $\delta_{Sim}(img_j, img_q) \rightarrow \infty$, então a imagem irá para o final do *ranking*. A ideia é aproximar, ainda mais, imagens similares à imagem-consulta e distanciar as não-similares.

Incluindo a interação com o usuário

Uma extensão do método proposto é integrar informações supervisionadas e, assim, agregar ainda mais semântica na consulta por similaridade. A proposta é usar o parecer do usuário, selecionando as imagens, dentro do conjunto de Imagens Dominantes, que são relevantes à sua necessidade.

A nova definição de similaridade proposta que permite a interação com o usuário é dado pela Definição 6.3.4.

Definição 6.3.4 *Dada uma imagem-consulta img_q e seu respectivo ranking $R_q = \{img_1, \dots, img_N\}$, um novo ranking $\tilde{R}_q$ é obtido considerando informação supervisionada e a distância de cada imagem $img_i \in R_q$ à sua respectiva Imagem Dominante $img_j \in D_{Im}$ através das seguintes condições:*

1. Se img_j foi selecionada pelo usuário como relevante, então

$$\delta_{SimD_{Im}}(img_i, img_q) = \delta_{Sim}(img_i, img_q) \quad (6.3)$$

2. Caso contrário:

$$\delta_{SimD_{Im}}(img_i, img_q) = \delta_{Sim}(img_i, img_q) + \delta_{Sim}(img_j, img_q) \quad (6.4)$$

A primeira condição da Definição 6.3.4 não altera a distância das imagens que estão sob influência de uma Imagem Dominante img_j selecionada como relevante pelo usuário. Assim, a similaridade é dada apenas pela distância da imagem à imagem-consulta. Caso contrário, sua distância é penalizada pela distância de sua imagem dominante à imagem-consulta.

6.4 Avaliações Experimentais

Esta seção apresenta um conjunto de experimentos realizados para verificar a eficácia do método proposto em várias bases de imagens, todas elas de domínio público e disponíveis na literatura. As mesmas citadas na seção 3.3.1 deste documento. São elas:

- Shapes: este é o conjunto de dados do MPEG-7 CE-Shape-1, composto por 1.400 imagens de formas de diferentes tipos, classificados em 70 classes com 20 exemplos em cada classe.
- Corel: consiste em 1.000 imagens formadas por cenas naturais complexas e cheias de detalhes, divididas em 10 classes, com 100 imagens em cada classe;

- Textura: consiste em 1.000 imagens, divididos em 40 amostras de 25 diferentes classes compostas por imagens de superfícies de materiais, tais como madeira, mármore e pele sob diferentes pontos de vista, escalas e condições de iluminação;
- Flowers: é a base de imagens da Oxford Flowers, composta por 1.360 imagens de 17 espécies diferentes de flores com 80 imagens por categorias.
- Médica: é um subconjunto da base do ImageCLEFmed 2007, composta por 5.042 imagens biomédicas de 32 categorias diferentes, tais como Raio-X, CT, MRI, etc;
- 15-Scenes: composta por 4.485 imagens de cenas naturais separadas em 15 categorias. Cada categoria é composta por 210 a 410 imagens;

A similaridade proposta ($SimD_{Im}$) tem duas variantes: *supervisionada* e *não-supervisionada*. Para uma análise comparativa, a versão $SimD_{Im}$ não-supervisionada foi comparada com a abordagem de ponderação baseada em TF-IDF (*Term Frequency-Inverse Document Frequency*), e a versão $SimD_{Im}$ supervisionada foi comparada com a técnica de realimentação de relevância QPM (*Query Point Movement*). Na avaliação supervisionada, a interação com o usuário foi realizada em apenas um ciclo de realimentação e utilizando o parecer do usuário limitado à apenas um conjunto de 10 imagens por consulta.

Para avaliar o desempenho da recuperação sobre todas as consultas de testes, a medida de avaliação adotada foi a média da precisão mAP (*mean Average Precision*). Tal medida de avaliação foi analisada usando três diferentes funções de similaridade δ_{Sim} : L_1 , L_2 e Canberra. O descritor utilizado para extrair as características das imagens foi o SIFT com a modelagem *Bag-of-Visual-Words* (BoVW), com 350 palavras-visuais. Para a base Shapes, o descritor utilizado foi o descritor *Bag-of-Saliency-Points* (BoSP).

Os gráficos da Figura 6.4 mostram os valores de mAP obtidos para cada base e para cada função de distância e os valores obtidos usando as abordagens de ponderação: TF-IDF, QPM, $SimD_{Im}$ Não-Supervisionada e $SimD_{Im}$ Supervisionada. Os resultados mostram que, para todas as bases, as abordagens de ponderação avaliadas permitiram aumentar a precisão do resultado e que as técnicas supervisionadas obtiveram valores maiores de precisão quando comparadas aos valores das técnicas não-supervisionadas.

A Tabela 6.1 mostra os ganhos obtidos utilizando cada uma das técnicas avaliadas quando comparadas à similaridade tradicional (*baseline*) baseada em análise par-a-par sem nenhuma informação contextual aplicada nessa análise. Considerando a avaliação não-supervisionada, o método proposto $SimD_{Im}$ obteve uma média de 4,3% de ganho de precisão nas consultas, enquanto que a técnica TF-IDF obteve apenas 1,5%. Já na avaliação supervisionada, o método proposto $SimD_{Im}$ obteve média de 46% de ganho, enquanto que a tradicional técnica de realimentação de relevância (QPM) obteve média de 21,2%. Os ganhos obtidos mostram que a similaridade proposta $SimD_{Im}$ apresenta uma

Figura 6.4: Valores de mAP (*mean Average Precision*) obtidos utilizando cada técnica de ponderação considerando diferentes funções de distâncias.

contribuição valiosa para aumentar o poder discriminatório na recuperação de imagens, aproximando ainda mais o resultado final ao esperado pelo usuário.

Com relação ao tempo computacional, a Tabela 6.2 mostra o tempo médio exigido em cada técnica para extrair informações (*mining*) e realizar a análise da similaridade

Base	Descritor	Distância	Ganho			
			Não-Supervisionado		Supervisionado	
			$SimD_{Im}$	TF-IDF	$SimD_{Im}$	QPM
Shape	BoSP	L_1	★ 7,0%	2,3%	46,8%	25,9%
		L_2	4,0%	1,8%	48,0%	25,2%
		Canberra	4,2%	2,1%	★ 56,9%	20,2%
Corel	BoVW	L_1	3,8%	0,6%	32,1%	14,6%
		L_2	★ 6,0%	2,9%	36,3%	18,7%
		Canberra	2,8%	1,5%	★ 37,0%	23,3%
Textura	BoVW	L_1	4,7%	2,1%	63,2%	28,0%
		L_2	★ 5,0%	1,4%	★ 64,1%	24,3%
		Canberra	3,5%	0,0%	56,1%	17,2%
Flower	BoVW	L_1	★ 6,9%	0,0%	49,4%	22,6%
		L_2	6,7%	1,7%	45,3%	20,9%
		Canberra	4,7%	2,0%	★ 50,0%	20,4%
Medica	BoVW	L_1	3,0%	0,9%	26,1%	10,6%
		L_2	★ 3,5%	2,3%	★ 29,1%	13,8%
		Canberra	1,8%	1,3%	27,3%	10,5%
15-Scenes	BoVW	L_1	1,8%	1,7%	46,3%	26,1%
		L_2	3,1%	1,3%	51,7%	32,1%
		Canberra	★ 5,2%	1,2%	★ 62,8%	26,6%

Tabela 6.1: Ganhos obtidos em cada técnica quando comparada ao método tradicional de similaridade baseada em análise par-a-par sem nenhuma ponderação contextual (*baseline*).

(*matching*). A técnica *baseline* não necessita extrair informações e executa a análise da similaridade em 5,66s. A técnica TF-IDF necessita de uma fase de *mining* para gerar o vetor de ponderação de atributos, essa tarefa exige 6,10s e é a técnica mais lenta para realizar o *matching*, pois exige mais operações do tipo *float* no cálculo da distância. A técnica QPM exige um tempo maior para calcular o novo vetor-de-características (*mining*), porém tem um tempo de *matching* próximo à *baseline*. As técnicas $SimD_{Im}$ supervisionada e não-supervisionada necessitam do mesmo tempo para gerar o conjunto de imagens dominantes, porém a técnica supervisionada é mais rápida que a técnica não-supervisionada, pois o custo de *matching* aumenta só para aquelas imagens selecionadas não-relevantes pelo usuário.

6.5 Considerações Finais

Neste capítulo foi apresentado um novo método de realizar consultas por similaridade baseado em conteúdo. A nova técnica apresentada, chamada de *Similaridade Baseada em Imagens Dominantes* ($SimD_{Im}$), enriquece a análise da similaridade levando em

Tabela 6.2: Tempo médio exigido em cada técnica para extrair informações (*mining*) e realizar a análise da similaridade (*matching*).

Técnica	Tempo Computacional [s]	
	<i>Mining</i>	<i>Matching</i>
Baseline	0	5,66
TF-IDF	6,10	7,22
QPM	8,13	5,76
$SimD_{Im}$ Não-Supervisionada	10,80	6,32
$SimD_{Im}$ Supervisionada		5,84

consideração a coleção de imagens do domínio da aplicação. A ideia é agregar informações contextuais no momento da realização da consulta, permitindo gerar um *ranking* de imagens que seja mais próximo da expectativa esperada pelo usuário.

A estratégia adotada pela técnica $SimD_{Im}$ é reduzir uma grande coleção de imagens em um sub-conjunto menor e mais diversificado de imagens, através do conceito de Imagens Dominadas, que analisa a similaridade entre as imagens a partir de uma imagem de referência. Dado um *ranking inicial* obtido por uma imagem-consulta, a primeira etapa da técnica $SimD_{Im}$ é obter um conjunto de Imagens Dominantes, composto por imagens que não são dominadas por nenhuma outra do próprio conjunto. A segunda etapa da técnica $SimD_{Im}$ é gerar um *novo ranking* re-calculando as distâncias das imagens usando o conjunto das Imagens Dominantes. A ideia é que se uma imagem dominante é similar à imagem-consulta, então todas as suas imagens dominadas também podem ser. Desse modo, é possível aproximar, ainda mais, as imagens similares e distanciar as não-similares.

Avaliações experimentais mostraram que a técnica $SimD_{Im}$ permite gerar um novo *ranking* com um ganho de precisão superior a 4,3% (média), porém para algumas bases esse ganho pode chegar até 7%. Esse é um ganho valioso se comparado com a técnica TF-IDF que, em média, só permite um ganho de 1,5% na precisão final dos resultados.

Uma versão supervisionada da técnica também foi apresentada. Essa nova versão permite a iteração do usuário com o sistema, em que o usuário seleciona as imagens relevantes dentro do conjunto de Imagens Dominantes. Essa iteração permite gerar um novo *ranking* com ganho médio de 46% na precisão final em apenas um ciclo de realimentação. A tradicional técnica de *Relevance Feedback* QPM obteve um ganho médio de apenas 21,2%.

Os ganhos obtidos mostram que o uso da similaridade proposta $SimD_{Im}$ permite quantificar de modo mais eficiente a similaridade entre imagens e, desse modo, aumentar o poder discriminatório na recuperação de imagens, reduzindo a lacuna semântica existente entre a percepção humana e a descrição computacional.

Conclusão

7.1 Considerações Finais

Para interpretar e representar o conteúdo de uma imagem é necessário um *descriptor* robusto capaz de: (i) extrair das imagens um número adequado e discriminatório de características; e (ii) uma função de distância eficaz e eficiente que compute a similaridade entre dois vetores-de-características. O fato é que uma imagem possui uma variedade enorme de dados a serem interpretados tanto quantitativa quanto qualitativamente, portanto uma imagem pode ser definida e tratada sob diferentes aspectos. Logo, aproximar um sistema de recuperação de imagens por conteúdo às expectativas do usuário final é uma tarefa desafiadora para a área de recuperação de informação, uma vez que o conceito de similaridade é algo subjetivo e depende de um contexto semântico esperado pelo usuário do sistema.

O objetivo final desse trabalho foi desenvolver técnicas para *caracterizar e medir a similaridade* de imagens, capazes de reduzir de maneira eficaz e eficiente o problema da descontinuidade semântica que ocorre entre a percepção humana e a descrição computacional, usando como motivação a abordagem baseada em dicionários-visuais. As diversas contribuições propostas neste trabalho de doutorado permitiram o avanço no estado-da-arte, aproximando a descrição e análise de imagens à subjetividade humana e desenvolvendo metodologias efetivas para caracterizar e quantificar a similaridade entre imagens.

7.2 Contribuições

Nesta tese foram propostos novos métodos de caracterização e análise de similaridade entre imagens para a área de CBIR, motivados pela metodologia *Bag-of-Visual-Words*. Resumidamente, as principais contribuições do presente trabalho para a área de CBIR são descritas a seguir:

- No Capítulo 3 foi apresentada uma extensa investigação da influência de diversos fatores na modelagem *Bag-of-Visual-Words* (BoVW), como: o tamanho do dicionário visual, o detector de pontos-de-interesse, o tipo de codificação (*soft* ou *hard*) e a função de distância utilizada para medir a similaridade. A constatação mais importante obtida nesses experimentos foi que todos os quatro fatores investigados são influentes para o desempenho da abordagem BoVW. O detector de pontos-de-interesse, por exemplo, é o fator com a maior influência na precisão final da recuperação de imagens. Além disso, a função de distância utilizada também interfere na precisão da recuperação, em geral a distância L_1 apresenta os melhores resultados com baixo custo computacional;
- O Capítulo 4 apresenta uma revisão bibliográfica de técnicas desenvolvidas na literatura que propõem incluir a disposição espacial das palavras-visuais de uma imagem na sua descrição final. Foram apresentadas também três novas técnicas de *pooling* que agregam a informação espacial das palavras-visuais na caracterização final da imagem, são elas: (i) *Weighted Histogram* (WH), (ii) *Bunch-of-2-grams* (Bo2g) e (iii) *Global Spatial Arrangement* (GSA). A principal vantagem dos três descritores é uma caracterização compacta e invariante à rotação. Cada um dos três descritores propostos tem sua particularidade de descrição e são propostas complementares que permitem sumarizar, de maneira mais descritiva, o conjunto das palavras-visuais de uma imagem e, assim, gerar uma caracterização final com maior poder discriminatório. O descritor Bo2g, por exemplo, mostrou-se com melhor desempenho para imagens de texturas, enquanto que GSA apresenta melhores resultados para imagens de cenas naturais e o descritor WH para imagens médicas. Quando comparados a um simples histograma de palavras-visuais, o descritor WH tem um ganho médio de 7% na recuperação de imagens, enquanto que os descritores GSA e Bo2g tem um ganho médio de 6%.
- No Capítulo 5 foi apresentado um novo descritor, chamado de *Bag-of-Saliency-Points* (BoSP), para a recuperação de imagens baseada em formas usando pontos salientes. O descritor proposto representa o conjunto dos pontos salientes de uma forma em um único vetor-de-características, análogo à abordagem *Bag-of-Visual-Words*. Esta estratégia permite calcular a similaridade

de formas usando funções de distância rápidas, como a L_1 . Também foi apresentado uma versão estendida do descritor BoSP que integra características globais na descrição final, referenciado como BoSP+GF. Esta versão estendida utiliza duas características globais para enriquecer a descrição semântica das formas, aprimorando de uma maneira eficiente a precisão na recuperação sem pagar computacionalmente mais por isso. A técnica proposta foi comparada com vários outros descritores da literatura e o principal ganho do descritor proposto é a alta qualidade no reconhecimento visual e uma rápida medida de similaridade, tornando o descritor BoSP uma contribuição valiosa na recuperação de imagens baseada em formas.

- O Capítulo 6 apresenta um novo método de realizar consultas por similaridade baseado em conteúdo, possibilitando a captura da semântica da consulta desejada pelo usuário. A proposta chamada de *Similaridade Baseada em Imagens Dominantes* ($SimD_{Im}$), enriquece a análise da similaridade agregando informações contextuais no momento da realização da consulta. A ideia é obter um conjunto de Imagens Dominantes, composto por imagens que não são dominadas por nenhuma outra do próprio conjunto, e usar essa informação para re-calcular as distâncias das imagens e, assim, gerar um *ranking* de imagens que seja mais próximo da expectativa esperada pelo usuário. Avaliações experimentais mostraram que a técnica $SimD_{Im}$ permite gerar um novo *ranking* com um ganho de precisão superior a 4,3% (média) usando a versão não-supervisionada do método. Já para a versão supervisionada, com a interação do usuário, o ganho médio é de 46% na precisão final em apenas um ciclo de realimentação.

7.3 Trabalhos Futuros

As contribuições apresentadas nesta tese geraram a necessidade de novos estudos, tanto para estender as técnicas desenvolvidas quanto para abordar outros fatores. A seguir são apresentadas algumas sugestões de tópicos que podem ser abordados em estudos futuros:

- Avaliar outras funções de ponderação na abordagem *Weighted Histogram*, bem como avaliar o impacto da codificação *soft* na recuperação de imagens;
- Investigar como a abordagem *Weighted Histogram* pode contribuir para identificar as regiões visuais mais importantes da imagem e empregar tal informação na recuperação de objetos;
- Avaliar o descritor BoSP com a codificação *soft* de palavras-visuais, bem como estender o descritor para formas em 3D;

- Combinar as abordagens WH, GSA e Bo2G em uma única caracterização final e avaliar ganho do poder discriminatório dessa combinação na recuperação de imagens;
- Propor e avaliar novos métodos de ponderação para computar uma nova similaridade considerando o conceito de Imagens Dominantes. A ideia é investigar novas estratégias para gerar um novo *ranking* considerando a similaridade de cada imagem à sua respectiva imagem dominante;
- Utilizar *modelos de atenção visual seletiva* nas etapas de extração de características e no cálculo da similaridade. Em outras palavras, a proposta futura é analisar como a similaridade entre imagens pode ser enriquecida com o apoio de suas informações visuais mais relevantes, ou seja, analisando o *contexto* visual da imagem. A estratégia é integrar contexto e conteúdo, a fim de obter um sistema de recuperação com desempenho perceptual mais condizente com o sistema visual humano.

7.4 Publicações Geradas

As publicações de artigos científicos que reflitam o grau de desenvolvimento do projeto e as contribuições para literatura foram realizadas de maneira continua durante todo o doutorado. A seguir são apresentas as publicações geradas até o presente momento:

1. PEDROSA, G.V.; TRAINA, A.J.M. From Bag-of-Visual-Words to Bag-of-Visual-Phrases Using n-Grams. In: *Conference on Graphics, Patterns and Images (SIBGRAPI)*, Arequipa Peru, p. 304-311, 2013.
2. PEDROSA, G.V.; TRAINA, A.J.M. Compact and Discriminative Approach for Encoding Spatial-Relationship of Visual Words. In: *ACM/SIGAPP Symposium On Applied Computing (SAC)*, Salamanca, Spain, p. 92-95, 2015.
3. PEDROSA, G.V.; TRAINA, A.J.M.; BARCELOS, C.A.Z. Shape Description Based on Bag-of-Saliency-Points. In: *ACM/SIGAPP Symposium On Applied Computing (SAC)*, Salamanca, Spain, p. 74-79, 2015.
4. BARBIERI, P.D.; PEDROSA, G.V.; TRAINA, A.J.M.; NOGUEIRA-BARBOSA, M.H. Vertebral Body Segmentation of Spine MR Images Using Superpixels. In: *IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, São Carlos, Brazil, p. 44-49, 2015.
5. PEDROSA, G.V.; RAHMAN, M.; ANTANI, S. K. ; DEMNER-FUSHMAN, D.; LONG, L. R.; TRAINA, A.J.M. Integrating Visual Words as Bunch of n-Grams for Effective Biomedical Image Classification. In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Steamboat Springs, USA, p. 431-436, 2014

6. PEDROSA, G.V.; TRAINA, A.J.M. ; TRAINA JR, C. Using sub-dictionaries for image representation based on the bag-of-visual-words approach. In: *IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, New York, USA, p. 165-168, 2014.
7. PEDROSA, G.V.; TRAINA, A.J.M. Making an Image Worth a Thousand Visual Words. *XI Workshop de Visão Computacional (WVC)*, p. 1-6, Out. 2015.
8. PEDROSA, G.V.; REZENDE, S.O.; TRAINA, A.J.M. Reducing the dimensionality of the SIFT descriptor and increasing its effectiveness and efficiency in image retrieval via bag-of-features. In: *Brazilian Symposium on Multimedia and the Web (WebMedia)*, São Paulo, Brazil, p. 139-142, 2012.

Referências Bibliográficas

- ANDALÓ, F.; MIRANDA, P.; DA S. TORRES, R.; FALCÃO, A. Shape feature extraction and description based on tensor scale. *Pattern Recognition*, v. 43, n. 1, p. 26 – 36, 2010.
- ARICA, N.; VURAL, F. T. Y. Bas: a perceptual shape descriptor based on the beam angle statistics. *Pattern Recognition Letters*, v. 24, n. 9-10, p. 1627 – 1639, 2003.
- ARULMOZHI, P.; ABIRAMI, S. Shape based image retrieval: A review. *Int. Journal on Comp. Science and Engineering*, v. 6, p. 14–153, 2014.
- AVILA, S.; THOME, N.; CORD, M.; VALLE, E.; DE A. ARAUJO, A. Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding*, v. 117, p. 453 – 465, 2013.
- AVNI, U.; GREENSPAN, H.; KONEN, E.; SHARON, M.; GOLDBERGER, J. X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words. *IEEE Trans. Med. Imaging*, v. 30, n. 3, p. 733–746, 2011.
- BAI, X.; RAO, C.; WANG, X. Shape vocabulary: A robust and efficient shape representation for shape matching. *IEEE Transactions on Image Processing*, v. 23, n. 9, p. 3935–3949, 2014.
- BATISTA, N. C.; LOPES, A. P. B.; ARAÚJO, A. D. A. Detecting buildings in historical photographs using bag-of-keypoints. In: *XXII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, 2009.
- BAY, H.; ESS, A.; TUYTELAARS, T.; VAN GOOL, L. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, v. 110, n. 3, p. 346–359, 2008.
- BOBER, M. Mpeg-7 visual shape descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 11, n. 6, p. 716–719, 2001.
- BOLOVINOU, A.; PRATIKAKIS, I.; PERANTONIS, S. Bag of spatio-visual words for context inference in scene classification. *Pattern Recognition*, v. 46, n. 3, p. 1039 – 1053, 2013.

- C. XU, J. L.; TANG, X. 2d shape matching by contour flexibility. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 31, p. 180186, 2009.
- CAO, Y.; WANG, C.; LI, Z.; ZHANG, L.; ZHANG, L. Spatial-bag-of-features. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, p. 3352–3359.
- CHANG, N. S.; FU, K. S. Query-by-pictorial-example. *IEEE Transactions on Software Engineering*, v. 6, n. 6, p. 519–524, 1980.
- CHAUDRY, Q.; RAZA, S. H.; YOUNG, A. N.; WANG, M. D. Automated renal cell carcinoma subtype classification using morphological, textural and wavelets based features. *Signal Processing Systems*, v. 55, n. 1-3, p. 15–23, 2009.
- CIACCIA, P.; PATELLA, M.; ZEZULA, P. M-tree: an efficient access method for similarity search in metric spaces. In *Proceedings of 23rd International Conference on Very Large Data Bases*, p. 426–435, 1997.
- CSURKA, G.; DANCE, C. R.; FAN, L.; WILLAMOWSKI, J.; BRAY, C. Visual categorization with bags of keypoints. In: *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, p. 1–22.
- DALIRI, M. R.; TORRE, V. Robust symbolic representation for shape recognition and retrieval. *Pattern Recognition*, v. 41, p. 17821798, 2008.
- DALITZ, C.; BRANDT, C.; GOEBBELS, S.; KOLANUS, D. Fourier descriptors for broken shapes. *EURASIP Journal on Advances in Signal Processing*, p. 161–161, 2013.
- DESELAERS, T.; KEYSERS, D.; NEY, H. Features for image retrieval: an experimental comparison. *Information Retrieval*, v. 11, n. 2, p. 77–107, 2008.
- F. R. SCHMIDT, E. T.; CREMERS, D. Efficient planar graph cuts with applications in computer vision. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 351356, 2009.
- FELZENSZWALB, P. F.; SCHWARTZ, J. D. Hierarchical matching of deformable shapes. *IEEE Conference on Computer Vision and Pattern Recognition (C)*, p. 1–8, 2007.
- FENG, J.; NI, B.; TIAN, Q.; YAN, S. Geometric lp-norm feature pooling for image classification. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, p. 2697–2704.
- FLICKNER, M.; SAWHNEY, H.; NIBLACK, W.; HUANG, Q.; ASHLEY, J.; DOM, B.; GORKANI, M.; HAFNER, J.; LEE, D.; PETKOVIC, D.; STEELE, D.; YANKER, P. Query by image and video content: the qbic system. *IEEE Computer*, v. 28, n. 9, p. 23–32, 1995.

- HARALICK, R. M.; SHANMUGAM, K.; DINSTEN, I. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, v. 3, n. 6, p. 610–621, 1973.
- HENNING MÜLLER, THOMAS DESELAERS, T. M. D. J. K. E. K. W. H. Overview of the imageclefmed 2007 medical retrieval and medical annotation tasks. *Advances in Multilingual and Multimodal Information Retrieval*, v. 1, p. 472 – 491, 2008.
- HOÀNG, N. V.; GOUET-BRUNET, V.; RUKOZ, M.; MANOUVRIER, M. Embedding spatial information into image content description for scene retrieval. *Pattern Recognition*, v. 43, n. 9, p. 3013–3024, 2010.
- HWANG, K.; LEE, H.; CHOI, D. Medical image retrieval: past and present. *Healthc Inform Res*, v. 18, n. 1, p. 3–9, 2012.
- J. WANG, X. BAI, X. Y. W. L.; LATECKI, L. J. Shape matching and classification using height functions. *Pattern Recognition Letters*, v. 33, p. 134143, 2012.
- JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *ACM Comput. Surv.*, v. 31, n. 3, p. 264–323, 1999.
- JEGOU, H.; HARZALLAH, H.; SCHMID, C. A contextual dissimilarity measure for accurate and efficient image search. In: *CVPR*, 2007.
- JEGOU, H.; PERRONNIN, F.; DOUZE, M.; SANCHEZ, J.; PEREZ, P.; SCHMID, C. Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.*, v. 34, n. 9, p. 1704–1716, 2012.
- JIANG, Y.-G.; NGO, C.-W.; YANG, J. Towards optimal bag-of-features for object categorization and semantic video retrieval. In: *International Conference on Image and Video Retrieval (CIVR)*, 2007, p. 494–501.
- JURIE, F.; TRIGGS, B. Creating efficient codebooks for visual recognition. In: *IEEE International Conference on Computer Vision (ICCV)*, 2005, p. 604–610.
- KE, Y.; SUKTHANKAR, R. Pca-sift: A more distinctive representation for local image descriptors. In: *CVPR (2)*, 2004, p. 506–513.
- KHAN, R.; BARAT, C.; MUSELET, D.; DUCOTTET, C. Spatial orientations of visual word pairs to improve bag-of-visual-words model. In: *British Machine Vision Conference, 3-7, 2012*, 2012, p. 1–11.
- KHOTANZAD, A.; HONG, Y. H. Invariant image recognition by zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 12, p. 489497, 1990.

- LAN, X.; ZITNICK, L.; SZELISKI, R. *Local bi-gram model for object recognition*. Relatório Técnico MSR-TR-2007-54, Microsoft Research, 2007.
- LATECKI, L. J.; LAKAMPER, R. Shape similarity measure based on correspondence of visual parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, p. 1185-1190, 2000.
- LAZEBNIK, S.; SCHMID, C.; PONCE, J. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 27, p. 1265-1278, 2005.
- LAZEBNIK, S.; SCHMID, C.; PONCE, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, p. 2169-2178.
- LIAO, S. X.; PAWLAK, M. On image analysis by moments. *IEEE Trans. Pattern Anal. Mach. Intell.*, v. 18, n. 3, p. 254-266, 1996.
- LINDBERG, T. Feature detection with automatic scale selection. *Int. J. Comput. Vision*, v. 30, n. 2, p. 79-116, 1998.
- LING, H.; JACOBS, D. W. Shape classification using the inner-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 29, p. 2862-2899, 2007.
- LOPES, A. P. B.; AVILA, S. E. F. D.; PEIXOTO, A. N. A.; OLIVEIRA, R. S.; COELHO, M. D. M.; ARAÚJO, A. D. A. Nude detection in video using bag-of-visual-features. In: *XXII Brazilian Symposium on Computer Graphics and Image Processing (SIB-GRAPI)*, 2009.
- LOWE, D. G. Object recognition from local scale-invariant features. In: *IEEE International Conference on Computer Vision (ICCV)*, 1999.
- LOWE, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, v. 60, n. 2, p. 91-110, 2004.
- LIU, M. B. C.; VEMURI, S.-I. A.; NIELSEN, F. Total bregman divergence and its applications to shape retrieval. *IEEE Conference on Computer Vision and Pattern Recognition. (CVPR)*, p. 3463-3468, 2010.
- LIU, M. B. C.; VEMURI, S.-I. A.; NIELSEN, F. Shape retrieval using hierarchical total bregman soft clustering. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, v. 34, p. 2047-2419, 2012.

- MANJUNATH, B. S.; MA, W. Y. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 18, n. 8, p. 837–842, 1996.
- MANSANO, A.; MATSUOKA, J.; AFONSO, L.; PAPA, J.; FARIA, F.; DA S TORRES, R. Improving image classification through descriptor combination. In: *Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2012, p. 324–329.
- MARÉE, R.; GEURTS, P.; PIATER, J. H.; WEHENKEL, L. Random subwindows for robust image classification. In: *CVPR*, IEEE Computer Society, 2005, p. 34–40.
- MARSZAEK, M.; SCHMID, C. Spatial weighting for bag-of-features. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, p. 2118–2125.
- MBANYA, E.; GERKE, S.; NDJIKI-NYA, P. Spatial codebooks for image categorization. In: *ACM International Conference on Multimedia Retrieval*, 2011, p. 50:1–50:7.
- MIKOLAJCZYK, K.; SCHMID, C. Scale and affine invariant interest point detectors. *Int. J. Comput. Vision*, v. 60, n. 1, p. 63–86, 2004.
Disponível em <http://dx.doi.org/10.1023/B:VISI.0000027790.02288.f2>
- MIKOLAJCZYK, K.; SCHMID, C. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, v. 27, n. 10, p. 1615–1630, 2005.
- MIKOLAJCZYK, K.; TUYTELAARS, T.; SCHMID, C.; ZISSERMAN, A.; MATAS, J.; SCHAFFALITZKY, F.; KADIR, T.; GOOL, L. V. A comparison of affine region detectors. *Int. J. Comput. Vision*, v. 65, n. 1-2, p. 43–72, 2005.
- MOKHTARIAN, F.; BOBER, M. *Curvature scale space representation: Theory, applications, and mpeg-7 standardization*. 1st ed. Springer Publishing Company, Incorporated, 2011.
- MOKHTARIAN, F.; SUOMELA, R. Robust image corner detection through curvature scale space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 20, p. 13761381, 1998.
- N. ALAJLAN, M. S. K.; FREEMAN, G. H. Geometry-based image retrieval in binary image databases. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 30, p. 10031013, 2008.
- NGUYEN, H.; PORIKLI, F. Support vector shape: A classifier based shape representation. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, v. 35, p. . 970982, 2013.

- NIEBLES, J. C.; WANG, H.; 0002, F.-F. L. Unsupervised learning of human action categories using spatial-temporal words. In: CHANTLER, M. J.; FISHER, R. B.; TRUCCO, E., eds. *BMVC*, British Machine Vision Association, 2006, p. 1249–1258.
- NILSBACK, M.-E.; ZISSERMAN, A. A visual vocabulary for flower classification. *Computer Vision and Pattern Recognition (CVPR)*, v. 1, p. 1447 – 1454, 2006.
- NOWAK, E.; JURIE, F.; TRIGGS, B. Sampling strategies for bag-of-features image classification. In: *European Conference on Computer Vision (ECCV)*, 2006.
- OHBUCHI, R.; OSADA, K.; FURUYA, T.; BANNO, T. Salient local visual features for shape-based 3d model retrieval. In: *IEEE Shape Modeling International*, 2008, p. 93–102.
- PAPA, J. P.; FALCÃO, A. X.; SUZUKI, C. T. N. Supervised pattern classification based on optimum-path forest. *International Journal of Imaging Systems and Technology*, v. 19, n. 2, p. 120–131, 2009.
- PEDROSA, G. V.; BARCELOS, C. A. Anisotropic diffusion for effective shape corner point detection. *Pattern Recognition Letters*, v. 31, n. 12, p. 1658–1664, 2010.
- PEDROSA, G. V.; BATISTA, M. A.; BARCELOS, C. A. Image feature descriptor based on shape salience points. *Neurocomputing*, v. 120, p. 156–163, 2013.
- PELLEG, D.; MOORE, A. X-means: Extending k-means with efficient estimation of the number of clusters. In: *International Conference on Machine Learning*, Morgan Kaufmann, 2000, p. 727–734.
- PENATTI, O. A.; SILVA, F. B.; VALLE, E.; GOUET-BRUNET, V.; TORRES, R. S. Visual word spatial arrangement for image retrieval and classification. *Pattern Recognition*, v. 47, n. 2, p. 705 – 720, 2014.
- PHILBIN, J.; ISARD, M.; SIVIC, J.; ZISSERMAN, A. Lost in quantization: Improving particular object retrieval in large scale image databases. In: *In CVPR*, 2008.
- R. GOPALAN, P. T.; CHELLAPPA, R. Articulation-invariant representation of non-planar shapes. *European Conference on Computer Vision (ECCV)*, p. 286299, 2010.
- RAHMAN, M. M.; ANTANI, S. K.; THOMA, G. R. Biomedical cbir using "bag of key-points" in a modified inverted index. In: *International Symposium on Computer-Based Medical Systems*, CBMS '11, 2011, p. 1–6 (*CBMS '11*,).
- ROCCHIO, J. *Relevance feedback in information retrieval*, cap. 14 Prentice-Hall Inc., p. 313–323, 1971.

- S. BELONGIE, J. M.; PUZICHA, J. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, p. 509522, 2002.
- S. BISWAS, G. A.; CHELLAPPA, R. An efficient and robust algorithm for shape indexing and retrieval. *IEEE Transactions on Multimedia*, v. 12, p. 372385, 2010.
- SANTOS, E. R. S.; LOPES, A. P. B.; VALLE, E. A.; ALMEIDA, J. M.; ARAÚJO, A. A. Vocabulários visuais para recuperação de informação multimídia. In: *Simpósio Brasileiro em Sistemas Multimídia e Web (WebMedia)*, 2010, p. 21–24.
- SAVARESE, S.; WINN, J.; CRIMINISI, A. Discriminative object class models of appearance and shape by correlatons. In: *IEEE Computer Vision and Pattern Recognition*, 2006, p. 2033–2040.
- SHAW, J. A.; FOX, E. A.; SHAW, J. A.; FOX, E. A. Combination of multiple searches. In: *The Second Text Retrieval Conference TREC-2*, 1994, p. 243–252.
- SILVA, F.; GOLDENSTEIN, S.; TABBONE, S.; DA S TORRES, R. Image classification based on bag of visual graphs. In: *IEEE International Conference on Image Processing*, 2013, p. 4312–4316.
- SIVIC, J.; RUSSELL, B.; EFROS, A.; ZISSERMAN, A.; FREEMAN, W. Discovering objects and their location in images. In: *IEEE International Conference on Computer Vision*, 2005, p. 370–377 Vol. 1.
- SIVIC, J.; ZISSERMAN, A. Video google: A text retrieval approach to object matching in videos. In: *International Conference on Computer Vision (ICCV)*, 2003, p. 1470–1477.
- SMITH, S. M.; BRADY, J. M. Susan - a new approach to low level image processing. *International Journal of Computer Vision*, v. 23, p. 45–78, 1995.
- SOARES, R. C.; SILVA, I. R.; GULIATO, D. Spatial locality weighting of features using saliency map with a bag-of-visual-words approach. In: *IEEE International Conference on Tools with Artificial Intelligence*, 2012, p. 1070–1075.
- STEINBACH, M.; KARYPIS, G.; KUMAR, V. A comparison of document clustering techniques. 2000a, p. 1–2.
- STEINBACH, M.; KARYPIS, G.; KUMAR, V. A comparison of document clustering techniques. In: *KDD Workshop on Text Mining*, 2000b, p. 1–20.
- SUN, W.; KISE, K.; CHAMPEIL, Y. Similar fragment retrieval of animations by a bag-of-features approach. In: *Proceedings of the 2012 10th IAPR International Workshop on Document Analysis Systems, DAS '12*, 2012, p. 140–144 (*DAS '12*).

- TAO, W.; ZHOU, Y.; LIU, L.; LI, K.; SUN, K.; ZHANG, Z. Spatial adjacent bag of features with multiple superpixels for object segmentation and classification. *Information Sciences*, v. 281, p. 373 – 385, 2014.
- TORRES, R.; FALCÃO, A.; DA F. COSTA, L. A graph-based approach for multiscale shape analysis. *Pattern Recognition*, v. 37, n. 6, p. 1163–1174, 2003.
- TORRES, R. D. S.; FALCÃO, A. X. Content-based image retrieval: Theory and applications. *Revista de Informática Teórica e Aplicada*, v. 13, p. 161–185, 2006.
- TORRES, R. S.; FALCÃO, A. X. Contour salience descriptors for effective image retrieval and analysis. *Image and Vision Computing*, v. 25, p. 3–13, 2007.
- TRAINA, J. C.; TRAINA, A. J. M.; FALOUTSOS, C.; SEEGER, B. Fast indexing and visualization of metric datasets using slim-trees. *IEEE Transactions on Knowledge and Data Engineering*, v. 14, n. 2, p. 244–260, 2002.
- VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, v. 3, n. 4, p. 209–235, 2010.
- WANG, Z.; LIANG, M. Locally affine invariant descriptors for shape matching and retrieval. *Signal Processing Letters*, v. 17, p. 803806, 2010.
- WILLETT, P.; EL-HAMDOUCHI, A. Comparison of hierarchic agglomerative clustering methods for document retrieval. *Computer Journal*, v. 32, n. 3, p. 220–227, 1989.
- WU, Z.; KE, Q.; ISARD, M.; SUN, J. Bundling features for large scale partial-duplicate web image search. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, p. 25–32.
- YANAI, K. Web image gathering with region-based bag-of-features and multiple instance learning. In: *ICME, IEEE*, 2009, p. 450–453.
- YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. In: *International Conference on Machine Learning (ICML)*, 1997, p. 412–420.
- YUAN, J.; WU, Y.; YANG, M. Discovery of collocation patterns: from visual words to visual phrases. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, p. 18–23.
- YUN, C.; SHIN, D.; JO, H.; YANG, J.; KIM, S. An experimental study on feature subset selection methods. In: *IEEE International Conference on Computer and Information Technology*, 2007, p. 77–82.

- Z. LI, W. QU, J. C. Z. S.; QI, H. Geometry-based 2d shape descriptor for retrieval in large database. *IEEE International Conference on Signal Processing*, v. 2, p. 10961101, 2012.
- ZHANG, E.; MAYO, M. Improving bag-of-words model with spatial information. In: *International Conference of Image and Vision Computing New Zealand*, 2010, p. 1–8.
- ZHANG, J.; MARSZALEK, M.; LAZEBNIK, S.; SCHMID, C. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, v. 73, n. 2, p. 213–238, 2007.
- ZHANG, L.; LIN, W. *Selective visual attention: Computational models and applications*. Wiley-IEEE Press, March 2013.
- ZHANG, Y.; JIA, Z.; CHEN, T. Image retrieval with geometry-preserving visual phrases. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, p. 809–816.