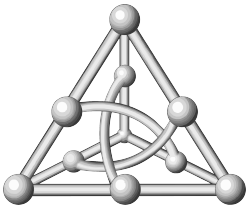
Informações Espaciais no Histograma de Palavras Visuais Usando Grafos

Jonatan Patrick Margarido Oruê
Orientador: Prof. Dr. Wesley Nunes Gonçalves



FACOM - Universidade Federal de Mato Grosso do Sul
Julho/2017

Informações Espaciais no Histograma de Palavras Visuais Usando Grafos[1]

Jonatan Patrick Margarido Oruê

Orientador: *Prof. Dr. Wesley Nunes Gonçalves*

Dissertação apresentada ao Programa de Mestrado *stricto sensu* em Ciência da Computação da Faculdade de Computação, mantido pela Universidade Federal do Mato Grosso do Sul, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação (Área de concentração: Visão Computacional).

UFMS - Campo Grande
Julho/2017

[1]Trabalho Realizado com Auxílio do CNPQ Proc. No: 133725/2015-4

Aos meus pais,
Cândido e Maria,

À minha irmã,
Joyce,

Ao meu sobrinho,
Leonardo.

Agradecimentos

À Deus por ter me dado saúde e força para superar todas as dificuldades.

Aos meus pais Cândido Ramão e Maria Suely, minha irmã Joyce e meu sobrinho Leonardo, pelo apoio, incentivo e amor incondicional.

Ao meu orientador Wesley Nunes Gonçalves, pelo suporte no pouco tempo que lhe coube, pelas suas correções, incentivos, ensinamentos e além disso, por sua amizade.

À todos os professores do curso, que foram tão importantes na minha vida acadêmica e no desenvolvimento desta dissertação.

Aos meus amigos de Ponta Porã e Campo Grande, companheiros de trabalhos e irmãos na amizade que fizeram parte da minha formação e que vão continuar presentes em minha vida com certeza.

Ao CNPq pela minha bolsa de mestrado e à Faculdade de Computação (FACOM-UFMS), pelo suporte e estrutura disponibilizados para o desenvolvimento de minha formação.

À todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

Resumo

ORUE, J. P. M. *Informações Espaciais no Histograma de Palavras Visuais Usando Grafos*. 2017. 53 p. Dissertação (Mestrado em Ciência da Computação) - Faculdade de Computação, Universidade Federal de Mato Grosso do Sul, Campo Grande, 2017.

Nos últimos anos, a classificação de imagens tem sido muito estudada na área de visão computacional por ser um tema desafiador. A classificação de imagens não é algo trivial de se executar, devido aos desafios impostos na captura das imagens (e.g., rotação, escala, iluminação, etc.). Aplicações de classificação de imagens incluem reconhecimento de cenas, objetos, faces, entre outras. Um dos métodos mais conhecidos na área de classificação de imagens é o Histograma de Palavras Visuais (BOVW). Primeiramente, o BOVW utiliza um método para localizar e descrever pontos de interesse das imagens. Em seguida, é utilizada uma técnica para o agrupamento dos pontos de interesse. O centroide de cada grupo é definido como uma palavra visual do vocabulário. A partir de uma nova imagem, cada ponto de interesse é rotulado em uma palavra visual do vocabulário. Por fim, o histograma da frequência de palavras visuais é utilizado como vetor de características para representar a imagem. Apesar dos resultados promissores, o BOVW não inclui informações espaciais em seu descritor final, incentivando pesquisas para analisar a influência dessas informações nos resultados de classificação de imagens. O método pirâmide espacial é a principal abordagem para inclusão de informações espaciais ao BOVW. Esse método consiste em após a rotulação dos pontos de interesse em palavras visuais, dividir a imagem em sub-regiões, construir um histograma de palavras visuais para cada sub-região e concatená-los para formar o descritor final da imagem. Este trabalho tem por objetivo acrescentar informações espaciais ao BOVW, combinando a técnica de pirâmide espacial com a modelagem das palavras visuais em grafos. Em nosso método, após a rotulação dos pontos de interesse em palavras visuais, a imagem é dividida pela pirâmide espacial. Em seguida, é construído um grafo para cada palavra visual em cada sub-região. Por fim, são extraídas medidas em relação a topologia dos grafos para formar o descritor final da imagem. Neste trabalho, foram feitos experimentos comparando o método proposto com os métodos BOVW e pirâmide espacial, o principal método para inclusão de informações espaciais presente na literatura. Além disso, foi apresentada uma análise de parâmetros do método proposto e uma comparação com os métodos presentes na literatura. Os experimentos foram realizados utilizando a base de imagens Caltech-101, devido a sua grande quantidade de imagens, suas variações intraclasse e as inúmeras variações em que as imagens foram capturadas. Os resultados apontaram que o método proposto apresenta um ganho na taxa de classificação correta média em relação aos métodos comparados. Por fim, o método proposto foi aplicado no problema do reconhecimento de vagas de estacionamento, em que o mesmo apresentou ótimos resultados.

Palavras-chaves: Classificação de Imagens, Histograma de Palavras Visuais, Topologia de Grafos.

Abstract

ORUE, J. P. M. *Spatial Information in Bag-of-Visual-Words Using Graphs*. 2017. 53 p. Dissertação (Mestrado em Ciência da Computação) - Faculdade de Computação, Universidade Federal de Mato Grosso do Sul, Campo Grande, 2017.

In recent years, image classification has been extensively studied in computer vision area to be a challenging topic. Classifying images is not something trivial to implement, due to the challenges in capturing the images (e.g., rotation, scale, lighting, etc.). Applications of image classification includes scene, object, face recognition, among others. One of the most popular methods in the area of image classification is the Bag-of-Visual-Words (BOVW). First, BOVW uses a method to locate and describe images keypoints. Then, a technique is used for the clustering of the keypoints. The centroid of each cluster is defined as one visual word of the vocabulary. From a new image, each keypoint is labeled in a visual word of the vocabulary. Finally, the histogram of the frequency of visual words is used as a feature vector to represent the image. Despite the promising results, the BOVW does not include spatial information in its final descriptor, encouraging research to analyze the influence of this information on image classification results. The spatial pyramid method is the main approach for inclusion of spatial information in the BOVW. This method consists of after labeling the keypoints in visual words, divide the image into sub-regions, build a histogram of visual words for each sub-region and concatenate them to form the final image descriptor. This work aims to add spatial information to BOVW by combining the spatial pyramid technique with the modeling of visual words in graphs. In our method, after labeling the keypoints in visual words, the image is divided by the spatial pyramid. Then a graph is built for each visual word in each sub-region. Finally, measures are taken with regard to the topology graph to form the final image descriptor. In this study, experiments were conducted comparing the proposed method with the methods BOVW and spatial pyramid, the main method for adding spatial information in the literature. Furthermore, it was presented a parameter analysis of the proposed method and compared with the methods in the literature. The experiments were performed using the Caltech-101, due to its large amount of images, their intra-class variance and the numerous variations in which the images were captured. The results showed that the proposed method has a gain in the average correct classification rate compared to previous methods. Finally, the proposed method was applied to the parking space recognition problem, in which it presented great results.

Key-words: Images Classification, Bag-of-Visual-Words, Graph Topology.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

SIFT *Scale Invariant Feature Transform*

BOVW *Bag-of-Visual-Words*

DoG *Difference of Gaussian*

PHOW *Pyramidal Histogram of Visual Words*

SURF *Speeded Up Robust Features*

HOG *Histograms of Oriented Gradients*

LBP *Local Binary Patterns*

LPC *Local Pairwise Codebook*

SFV *Spatial Fisher Vectors*

MoG *Mixture of Gaussians*

FV *Fisher Vector*

MP Método Proposto

TCCM Taxa de Classificação Correta Média

MF Medida-f

PR Precisão

RE Revocação

SVM *Support Vector Machines*

FN Falso Negativo

FP Falso Positivo

VN Verdadeiro Negativo

VP Verdadeiro Positivo

NC Número de Características

NI Número de Imagens

CAPÍTULO

1

Introdução

Neste Capítulo é apresentada uma descrição desta dissertação, com o objetivo de fornecer uma visão geral dos problemas tratados e dos objetivos principais do trabalho de pesquisa. O Capítulo está organizado da seguinte maneira: na Seção 1.1 é apresentada a contextualização e motivação sobre o tema de pesquisa tratado nesta dissertação, além das principais contribuições do trabalho; na Seção 1.2 é apresentada a revisão da literatura realizada para a execução deste trabalho; na Seção 1.3 são apresentados os objetivos gerais e específicos do trabalho; por fim, na Seção 1.4 é apresentada a organização da dissertação, com uma descrição resumida do conteúdo abordado em cada capítulo.

1.1 Contextualização e Motivação

A classificação de imagens tem como um de seus objetivos, determinar a presença de objetos em imagens, ou reconhecê-las como um tipo de cena em particular (e.g., praia, montanha ou cidade) (Csurka et al., 2004). O reconhecimento de cenas e objetos é, atualmente, um dos problemas mais estudados em visão computacional, pelo seu elevado grau de dificuldade. Essas dificuldades ocorrem devido as inúmeras transformações em que as imagens estão sujeitas (e.g., rotação, escala, iluminação, etc.) (Brown and Susstrunk, 2011). Em geral, os métodos para reconhecimento de cenas e objetos são divididos em duas etapas, detecção de pontos de interesse e, posteriormente, utilização de um descritor capaz de gerar atributos para representar os pontos de interesse (Liu et al., 2016).

Os métodos de reconhecimento de cenas e objetos em estado da arte utilizam a abordagem de histograma de palavras visuais (do inglês - *Bag-of-Visual-Words* – BOVW) proposta por Csurka et al. (2004) ou redes neurais convolucionais (Shah et al., 2016). O BOVW consiste em construir um vocabulário de palavras visuais a partir de pontos de interesse detectados por descritores locais. A primeira etapa para a construção do vocabulário é a detecção e descrição de pontos de interesse. Existem hoje na literatura diversos algoritmos que tem como objetivo detectar e/ou descrever pontos. O SIFT (Transformada de Características Invariante a Escala, do inglês - *Scale Invariant Feature Transform*) foi proposto por Lowe (2004) e é um dos descritores mais utilizados atualmente por ter apresentado ótimos resultados em aplicações de reconhecimento de cenas e objetos (Brown and Susstrunk, 2011). Além do SIFT, podemos

citar o SIFT Denso (Liu et al., 2016), *Speeded-UP Robust Features* – SURF (Bay et al., 2006), *Pyramidal Histogram of Visual Words* – PHOW (Bosch et al., 2007), *Histograms of Oriented Gradients* – HOG (Dalal and Triggs, 2005) e o *Local Binary Patterns* – LBP (Ojala et al., 2002).

Após obter os pontos de interesse das imagens de treinamento, os mesmos são concatenados em um conjunto de descritores locais. Então, o BOVW utiliza um método para separar os descritores locais em grupos. Normalmente o método usado é o *k-means*, por produzir resultados simples e intuitivos para serem interpretados, além de ter complexidade linear (Wu et al., 2008). A partir do agrupamento dos descritores locais, o centroide de cada grupo é definido como uma palavra visual. Dessa forma, são extraídos k centroides que formam o vocabulário de palavras visuais. Após a construção do vocabulário, cada ponto de interesse localizado em uma determinada imagem é rotulado em uma palavra do vocabulário de palavras visuais e um histograma da frequência de palavras visuais é construído para formar o vetor de características que descreve a imagem.

Apesar dos resultados promissores, o método BOVW não inclui informações espaciais em seu descritor final, gerando incentivo a novas pesquisas para averiguar a importância da inclusão de informação espacial nas características extraídas. Para compreender melhor a deficiência espacial do BOVW, considere as duas imagens sintéticas da Figura 1.1. Nestas imagens, cada pixel foi rotulado em uma palavra visual, a partir de um vocabulário de palavras visuais com $k = 3$. Como pode ser observado, os histogramas de palavras visuais de ambas as imagens são iguais, pois a frequência das 3 palavras visuais são iguais nas duas imagens. Dessa forma, o BOVW não considera a localização espacial das palavras visuais na imagem.

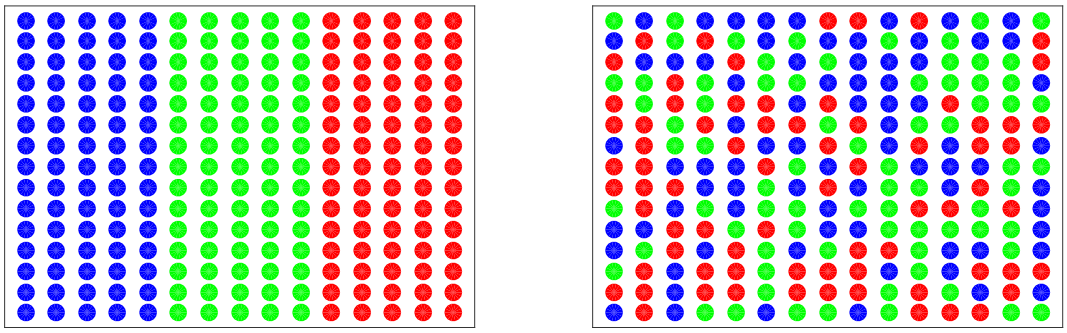


Figura 1.1: Exemplo de imagens com seus respectivos pontos de interesse rotulados por palavras do vocabulário visual. Na primeira imagem (a), as palavras visuais estão distribuídas uniformemente, enquanto na segunda (b) as palavras visuais estão distribuídas de forma aleatória.

O método de pirâmide espacial proposto por Lazebnik et al. (2006), é a principal proposta para a inclusão de informações espaciais 2D ao BOVW. Os primeiros passos do pirâmide espacial são os mesmos do BOVW (construção do vocabulário de palavras visuais e rotulação dos pontos de interesse em palavras visuais). Feito isso, o método divide a imagem em sub-regiões menores calculando para cada sub-região um histograma de palavras visuais e, em seguida, eles são concatenados para formar o descritor final da imagem. A Figura 1.2, ilustra um exemplo em que duas imagens sintéticas diferentes passam pela última etapa do pirâmide espacial, ou seja, a divisão em sub-regiões. No exemplo, podemos perceber que aplicando o método pirâmide espacial, os descritores finais são diferentes. O pirâmide espacial é uma alternativa que tem apresentado resultados promissores em métodos de reconhecimento de objetos (Lazebnik et al., 2006). A informação espacial fornecida por esse método é suficiente

para aumentar a taxa de classificação correta média (Lazebnik et al., 2006).

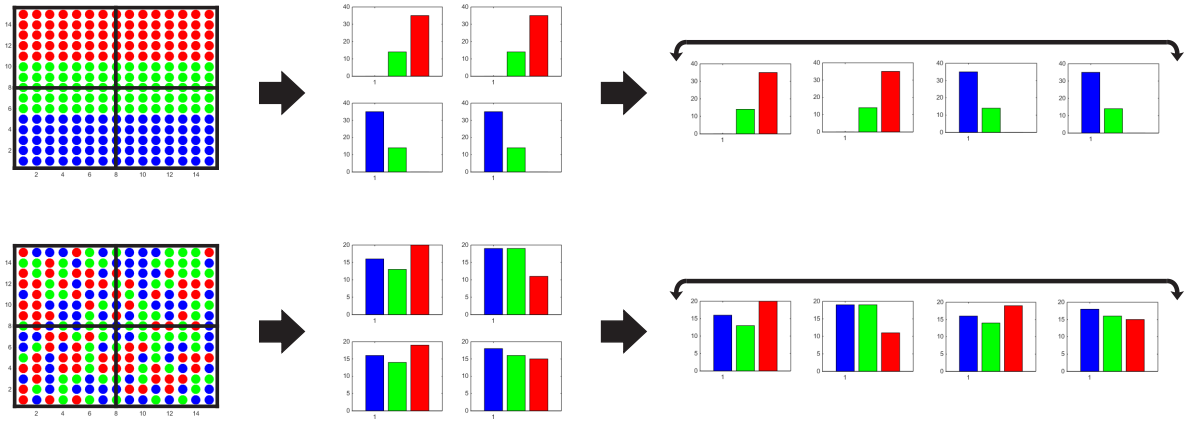


Figura 1.2: Exemplo de duas imagens sintéticas com seus respectivos pontos de interesse rotulados por palavras do vocabulário visual. Ambas estão usando o método pirâmide espacial para gerar o descritor final. Desse modo, a imagem é sub-dividia em regiões e, em seguida, é computado o histograma de cada região. Por fim, os histogramas são concatenados para formar o descritor final da imagem. Com o exemplo, podemos perceber que utilizando o método pirâmide espacial, os descritores finais para as imagens são diferentes por considerar a localização espacial dos pontos de interesse.

Contudo, caso a frequência de palavras visuais estiver igualmente distribuídas em cada sub-região da Figura 1.2, o descritor extraído pelo método pirâmide espacial será o mesmo em ambas as imagens. Na atualidade, a utilização de grafos em visão computacional tem apresentado resultados promissores, principalmente na área de análise de texturas (Gonçalves, 2013). Com isso, nossa proposta consiste em criar um novo método para descrever as palavras visuais obtidas pelo BOVW, combinando o método de pirâmide espacial com grafos. O método proposto se baseia em construir o vocabulário de palavras visuais e rotular os pontos de interesse em palavras visuais (duas primeiras etapas do BOVW). Feito isso, é aplicada a técnica de pirâmide espacial para dividir a imagem em regiões cada vez menores. Então, em cada sub-região da imagem é modelado um grafo para cada palavra visual, ou seja, são modelados k grafos para cada sub-região. Após a modelagem dos k grafos, é extraído o grau médio para cada grafo, explorando assim mais informações espaciais do método, a partir da topologia do grafo. Posteriormente, os graus médios dos grafos são concatenados, formando um vetor de tamanho k em cada sub-região da imagem. Por fim, os vetores extraídos de cada sub-região em um determinado nível da pirâmide e os vetores extraídos das pirâmides de níveis inferiores são concatenados, formando assim o descritor final da imagem.

A base de imagens Caltech-101 foi utilizada nos experimentos por ser uma das mais utilizadas atualmente (Rosa et al., 2016; Han et al., 2017). Os descritores locais SIFT Denso, PHOW e PHOW-RGB foram utilizados para detectar e descrever os pontos de interesse das imagens. Com o propósito de se encontrar o melhor conjunto de parâmetros para utilização do método proposto, foi feita uma análise de parâmetros envolvendo os descritores locais, o tamanho do vocabulário de palavras visuais e o raio utilizado na modelagem dos grafos. Para validar os resultados do método proposto, o mesmo foi comparado com os métodos BOVW e pirâmide espacial, onde foi apontado uma superioridade do método proposto em relação aos métodos comparados. Através dos experimentos realizados por este trabalho, o método proposto obteve uma taxa de classificação correta média de 80.95%, enquanto os métodos BOVW e pirâmide espacial alcançaram 70.66% e 80.71%, respectivamente.

Para validar o método proposto em um problema real, o mesmo foi aplicado no reconhecimento de vagas de estacionamento (Falcão et al., 2013). A base de imagens *PKLot* foi utilizada

nos experimentos, por conter exemplos de estacionamentos reais, com imagens capturadas sob diferentes ângulos de visão e condições climáticas variadas (e.g., nublados, ensolados, chuvosos) (De Almeida et al., 2015). Com os resultados, é possível perceber que o método proposto se mostrou robusto na solução do problema do reconhecimento de vagas de estacionamento, onde o mesmo alcançou uma taxa de classificação correta média de 99.81%.

1.2 Revisão da literatura

Nesta seção, são apresentados os principais trabalhos que estenderam o histograma de palavras visuais para o reconhecimento de cenas e objetos. Como mostrado na Seção 1.1, o método BOVW por apenas considerar a frequência de palavras visuais em uma imagem, acaba não extraindo informações em relação a posição espacial dos pontos de interesse. Desse modo, diversas pesquisas foram realizadas com o intuito de avaliar a importância da inclusão de informações espaciais ao BOVW. A principal abordagem para inclusão de informações espaciais ao BOVW, é o pirâmide espacial Lazebnik et al. (2006). Desse modo, muitas abordagens utilizaram esse método como base para construção de outras abordagens. A seguir, são apresentados alguns trabalhos que propuseram a inclusão de informações espaciais ao BOVW.

(a) Abordagens baseadas em correlogramas: Um método que incorpora o conjunto de informações de aparência e forma para reconhecimento de objetos foi proposto por Savarese et al. (2006). Nesse trabalho, foi introduzido o conceito de correlogramas de palavras visuais para obtenção de informações espaciais. O método de correlogramas visa analisar as correlações de pares de palavras visuais em uma determinada localização espacial. Com isso, é possível calcular como a correlação das palavras visuais mudam conforme a distância. Desse modo, é possível calcular informações locais e globais, dependendo do tamanho do raio de abrangência. O método mostrou ser robusto para transformações geométricas, oclusões e falta de informações, porém é totalmente dependente do raio de abrangência para construir as correlações de palavras visuais.

A abordagem proposta por Quack et al. (2007) é baseada em técnicas de mineração de dados, tais como *Support Vector Machine* (SVM) e pode encontrar características frequentes entre dezenas de milhares de candidatos em questão de segundos. Com base nas configurações mineradas, foi desenvolvido um método para selecionar as características que têm alta probabilidade de não pertencerem ao objeto. A técnica é concebida como uma camada de transformação intermediaria para filtrar a grande quantidade de características desordenadas devolvidas pelo estágio de extração para facilitar as tarefas de processamento de alto nível, tais como detecção de objetos.

Liu et al. (2008) propuseram um método baseado no BOVW que realiza a extração e seleção de características de forma integrada. As características espaciais de ordem superior (histograma espacial) são progressivamente extraídas com base em características de ordem inferior (BOVW), evitando assim uma computação exaustiva. O método pode ser baseado em qualquer algoritmo de seleção e extração de características. No método proposto por Liu et al., a ideia é executar ambos métodos simultaneamente. Os resultados experimentais mostraram que o método é computacionalmente mais eficiente do que as abordagens anteriores, sem perda de precisão.

O método proposto por Morioka and Satoh (2010) faz o emparelhamento de características locais, chamado *Local Pairwise Codebook* (LPC). Primeiramente é concatenado cada par de descritores espacialmente próximos, e são tratados como descritores em um espaço de características. O tamanho do LPC é controlado diretamente por um algoritmo

de agrupamento, atribuindo a cada par de descritores, o grupo mais próximo. A complexidade de tempo cresce conforme o número de pares formados. Assim, para reduzir significativamente a complexidade, foi proposto uma técnica de atribuição de agrupamento emparelhadas eficiente.

(b) Abordagens baseadas em pirâmides espaciais: Krapac et al. (2011) mostraram que utilizar o núcleo espacial de Fisher (do inglês - *Spatial Fisher Vectors* - SFV) possui um excelente desempenho quando usado com classificadores lineares. O SFV era utilizado somente para codificar informações de aparência. Nesse trabalho, as palavras visuais foram modeladas utilizando Misturas de Gaussianas (do inglês - *Mixture of Gaussians* - MoG) e em seguida, computado o núcleo de Fisher para esses modelos. Essa abordagem apresentou uma maior acurácia em relação ao método de pirâmides espaciais.

O objetivo do trabalho proposto por Sánchez et al. (2012) foi incluir informações sobre a disposição espacial em assinaturas de imagens com base em estatísticas médias. Foram propostas alternativas de pirâmides espaciais para o reconhecimento de objetos. O trabalho foi concentrado no FV (*Fisher Vector*), no entanto, o trabalho poderia ser expandido a outras técnicas baseadas no BOVW. Nesse trabalho, foram propostas duas formas diferentes e complementares para incluir a informação espacial na assinatura de imagem que têm como alvo duas fontes de variação. A primeira é devido ao fato do FV ser calculado a partir de um conjunto finito de descritores. A segunda vem do fato de que a proporção do objeto pode variar entre duas imagens da mesma classe. Reduzir estas fontes de variação, aumentaria a separabilidade linear e, portanto, a precisão da classificação melhoraria.

No artigo proposto por Bolovinou et al. (2013), é apresentada uma nova abordagem para adicionar contexto na representação das imagens, através de informações espaciais ordenadas de palavras visuais. O método proposto introduz um histograma espacial de palavras visuais (BOSVW) obtido por agrupamento de conjuntos de correlação de palavras visuais. Especificamente, o algoritmo de agrupamento *k-means* é utilizado, representando a grande dimensionalidade e a dispersão dos descritores propostos. Os resultados experimentais em quatro conjuntos de dados mostraram que o método proposto melhora significativamente o BOVW e se compara favoravelmente às abordagens de classificação de cenas baseadas em contexto existentes.

No trabalho proposto por Khan et al. (2015), foi apresentado uma nova maneira de incluir informações de distância e ângulo na representação do BOVW. O método fornece uma representação computacionalmente eficiente, adicionando informações espaciais relativas a paridade de palavras visuais com base na distância dos pontos de interesse. Experimentos em bases de imagens desafiadoras demonstram que esse método supera ou é competitivo com os concorrentes. Foi mostrado também, que esse método fornece informações complementares importantes para a correspondência da pirâmide espacial e pode melhorar o desempenho geral.

1.3 Objetivos

1.3.1 Objetivo Geral

Este trabalho tem como objetivo propor um método para incluir informações espaciais ao histograma de palavras visuais - BOVW, combinando o método pirâmide espacial com informações extraídas de grafos. As informações espaciais são incluídas por meio da modelagem de k grafos em cada sub-região dividida pelo pirâmide espacial, e assim, extrair informações por meio da topologia dos grafos para compor o descritor final da imagem.

1.3.2 Objetivos Específicos

Os objetivos específicos podem ser divididos da seguinte forma:

1. Estudar, compreender e implementar os algoritmos:
   a *Scale Invariant Feature Transform* - SIFT;
   b SIFT Denso;
   c *Pyramidal Histogram of Visual Words* - PHOW;
   d PHOW-RGB;
   e Histograma de Palavras Visuais - BOVW;
   f Pirâmide Espacial.
2. Levantar os bancos de imagens a serem utilizados.
3. Implementar e validar um método baseado em extrair características a partir da descrição da topologia de grafos.
4. Comparar o método proposto com algoritmos em estado da arte de reconhecimento de cenas e objetos.
5. Registar e divulgar os resultados obtidos.

1.4 Organização do Texto

Este trabalho está organizado da seguinte maneira:

- **Capítulo 2:** neste capítulo, são apresentados os referenciais teóricos pertencentes aos métodos utilizados neste trabalho, como os descritores locais SIFT Denso, PHOW e PHOW-RGB, o método de extração de características *Bag-of-Visual-Words*, o principal método para inclusão de informações espaciais ao BOVW (pirâmide espacial) e uma breve introdução aos grafos.
- **Capítulo 3:** o método proposto é detalhado neste capítulo, o qual foi baseado na modelagem de grafo e extração de medidas da topologia do grafo. Além disso, são apresentadas as invariâncias do método e equivalência do método proposto com os métodos BOVW e Pirâmide Espacial.
- **Capítulo 4:** o banco de imagens utilizado, os protocolos usados nos experimentos, a análise e comparação dos resultados alcançados são descritos neste capítulo.
- **Capítulo 5:** neste capítulo, são apresentados resultados obtidos da aplicação do método proposto no problema do reconhecimento de vagas de estacionamento.
- **Capítulo 6:** finalmente, neste capítulo são discutidas as conclusões dos resultados, as limitações da proposta e os trabalhos futuros.
- **Apêndice A:** neste apêndice, são apresentados resultados complementares dos experimentos realizados.

CAPÍTULO

2

Referencial Teórico

Neste capítulo, são apresentados os referenciais teóricos para compreensão deste trabalho. A Seção 2.1 descreve em detalhes os descritores locais SIFT, SIFT Denso e PHOW, que tem como objetivo detectar e descrever pontos de interesse na imagem, afim de caracterizá-la. A Seção 2.2 descreve o histograma de palavras visuais - BOVW. A técnica para inclusão de informações espaciais ao método BOVW, conhecida como pirâmide espacial é apresentada na Seção 2.3. Por fim, a Seção 2.4 apresenta uma introdução aos grafos que são utilizados na proposta desse trabalho.

2.1 Descritores Locais

2.1.1 Scale Invariant Feature Transform - SIFT

O SIFT (Transformada de Características Invariante a Escala, do inglês - *Scale Invariant Feature Transform*) foi proposto por Lowe (2004) e é, atualmente, um dos mais importantes descritores de características locais, tendo apresentado resultados promissores em aplicações de reconhecimento de cenas e objetos (Brown and Susstrunk, 2011). As características extraídas pelo SIFT são invariantes à escala e rotação, além de serem parcialmente invariantes a mudanças de iluminação e ponto de vista. O SIFT pode ser dividido em 2 etapas principais: i) detecção e ii) descrição de pontos de interesse. Os pontos de interesse são localizados tanto no domínio espacial quanto na frequência, reduzindo assim a influência negativa de ruídos, oclusões e desordem (Lowe, 2004). As duas etapas principais do SIFT podem ser divididas em outras quatro etapas: i) construção do espaço de escalas, ii) localização de pontos de interesse, iii) atribuição de uma orientação e a iv) extração de descritores para cada ponto. As quatro etapas do SIFT são detalhadas a seguir.

Construção do espaço de escalas

Para tornar o método invariante a escala, a primeira etapa do SIFT consiste em convoluir uma imagem $I(x,y)$ em diferentes escalas por meio do filtro Gaussiano com diferentes núcleos,

formando o conjunto de imagens L, conforme a Equação 2.1.

$$L(x,y,\sigma) = I(x,y) \otimes G(x,y,\sigma) \tag{2.1}$$

onde (x,y) representa a localização espacial dos pixels da imagem, $G(x,y,\sigma)$ é o núcleo Gaussiano $G(x,y,\sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{\sigma^2}}$ com desvio padrão σ e o símbolo $\otimes$ é a operação de convolução.

Para tornar o método tolerante a escalas diferentes, Lowe considera que é necessário convoluir a imagem de uma escala inicial σ_0 até $2\sigma_0$. Portanto, para se gerar s intervalos, o fator de incremento k é definido como $k = 2^{1/s}$. Sendo assim, as escalas (desvio padrão do núcleo Gaussiano) correspondem a $[\sigma_0, k\sigma_0, 2k\sigma_0, ...]$.

As regiões interessantes da imagem estão localizadas em regiões onde há mudança de intensidade, isto é, onde a derivada é alta. Deste modo, a diferença de Gaussianas – DoG (Vedaldi and Fulkerson, 2010) é utilizada por ser uma forma eficiente de encontrar pontos de interesse com mudanças de intensidade em diferentes escalas. A DoG é calculada pela diferença entre imagens do conjunto L de duas escalas vizinhas, separadas por uma constante k, conforme a Equação 2.2. Após obter o conjunto de imagens L convoluídas por filtros Gaussianos, a técnica DoG (Equação 2.2) é aplicada como podemos ver na Figura 2.1, gerando um novo conjunto de imagens D formadas pela diferença de Gaussianas.

$$\begin{aligned} D(x,y,\sigma) &= (G(x,y,k\sigma) - G(x,y,\sigma)) \otimes I(x,y) && (2.2)\\ &= G(x,y,k\sigma) \otimes I(x,y) - G(x,y,\sigma) \otimes I(x,y) && (2.3)\\ &= L(x,y,k\sigma) - L(x,y,\sigma). && (2.4) \end{aligned}$$

As imagens do conjunto D resultam da diferença entre as imagens de escalas vizinhas que são submetidas ao filtro Gaussiano com as escalas σ e $k\sigma$. Os pixels com valores altos em $|D(x,y,\sigma)|$ correspondem a lugares onde há mudança brusca de intensidade. Um dos motivos em se utilizar essa função é o fato de sua complexidade computacional ser baixa, pois somente é utilizada subtração de imagens convoluídas (Lowe, 1999).

Ao final da DoG, é gerado o que é chamado de uma oitava, na qual obtêm-se os conjuntos de imagens L e D para a imagem analisada. Este processo é repetido para um número desejado de oitavas. Na primeira oitava a imagem original tem seu tamanho dobrado, a segunda oitava é a imagem original, a partir da terceira oitava a imagem tem seu tamanho reduzido pela metade a cada iteração. Em cada oitava a imagem é convoluída por filtros Gaussianos em diferentes escalas, criando um conjunto L de $s+3$ imagens convoluídas por oitava de maneira que a detecção de extremos cubra toda oitava.

Na Figura 2.1, mais a esquerda, temos a imagem de entrada de cada oitava (perceba que as resoluções são diferentes). Ao centro, temos a representação da imagem de entrada convoluída por filtros Gaussianos em diferentes escalas. No exemplo, foram produzidas 5 imagens resultantes da aplicação dos filtros Gaussianos. A direita temos as imagens de saída, que representam o resultado da aplicação da técnica DoG, esse processo é denominado uma oitava. Ao fim de uma oitava, a imagem de entrada é redimensionada pela metade e o processo é realizado novamente.

Localização de pontos de interesse

Uma vez obtido o conjunto de imagens D resultante da diferença de Gaussianas em cada oitava, é aplicado o detector de extremos (máximos e mínimos) locais em $D(x,y,\sigma)$ a fim de identificar pontos de interesse na imagem. Cada ponto (x,y) de $D(x,y,\sigma)$ é comparado com seus oito vizinhos em sua escala, nove vizinhos da escala anterior e os nove vizinhos da escala

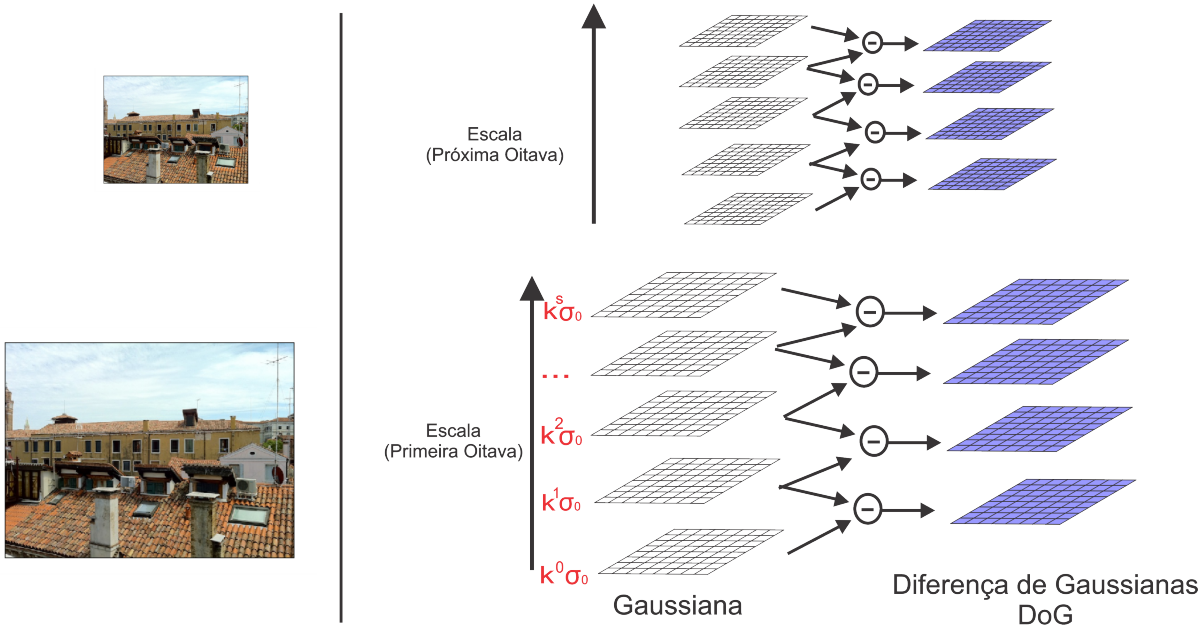


Figura 2.1: Para cada oitava do espaço de escalas, a imagem inicial é convoluída repetidamente por filtros Gaussianos para produzir o conjunto de imagens convoluídas L. As imagens convoluídas por Gaussianas adjacentes são subtraídas para produzir o conjunto de imagens D que representam a diferença de Gaussianas (imagens à direita). Para cada oitava, a imagem de entrada é reduzida por um fator de 2, e o processo é repetido.

posterior, totalizando vinte e seis elementos. A Figura 2.2 ilustra como é feito o procedimento, neste caso, o pixel marcado em vermelho é o pixel candidato a ponto de interesse e os demais pixels marcados em azul são comparados com o pixel candidato. Se $D(x,y,\sigma)$ tiver um valor maior ou menor que todos os 26 elementos comparados com ele, então, (x,y) é considerado um ponto de interesse. O custo desta seleção é razoavelmente baixo, pois a maioria dos pontos são eliminados após as primeiras verificações (Brown and Lowe, 2002).

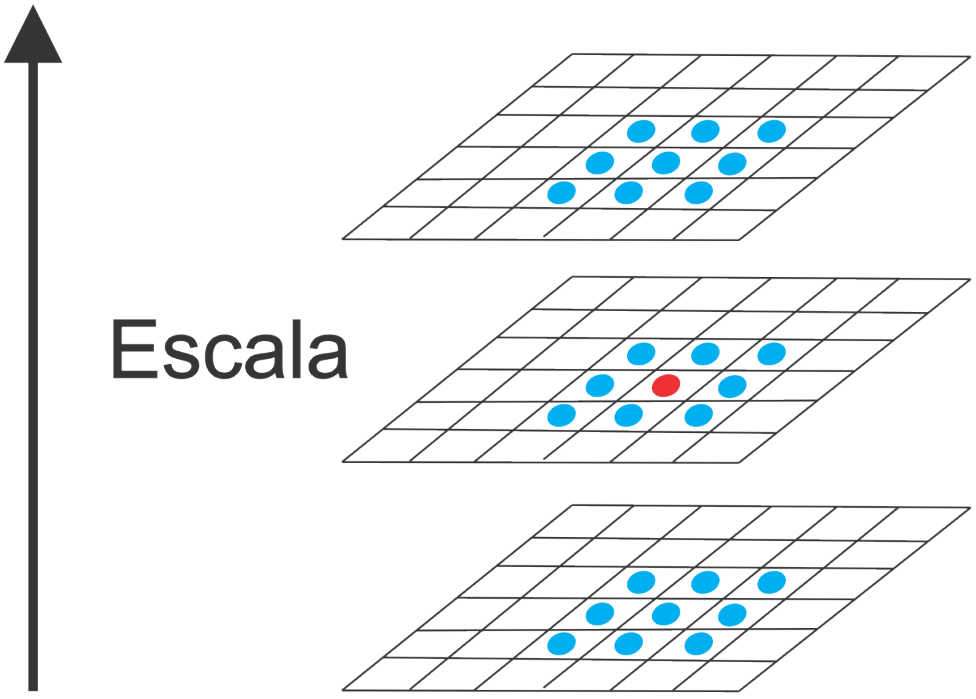


Figura 2.2: Exemplo da detecção de máximos e mínimos entre imagens resultantes da diferença de Gaussianas. Neste exemplo, o pixel marcado em vermelho é comparado em uma região 3×3 com os 8 vizinhos na escala atual e 18 pixels das escalas adjacentes.

Após realizar a detecção apresentada, obtêm-se pontos de interesse que ainda não são confiáveis devido a ruídos ou pequenas variações, dessa forma, são aplicadas duas etapas de eliminação de pontos. Na primeira etapa, os valores de $D(x,y,\sigma)$ são normalizados entre $[0,1]$, e então, são eliminados os pontos de baixo contraste cujo $D(x,y,\sigma) < 0.03$ (esse limiar é recomendado por Lowe). A seguir, são eliminados os pontos com resposta de borda forte em uma única direção, ou seja, pontos instáveis em relação ao ruído. Esses pontos são encontrados através de um procedimento baseado na matriz Hessiana, que é formada pelas derivadas parciais D do pixel candidato a ponto de interesse, como mostrado na Equação 2.5.

$$H = \begin{vmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{vmatrix} \tag{2.5}$$

Os pontos são rejeitados conforme a seguinte condição:

$$Tr(H) = D_{xx} + D_{yy} \tag{2.6}$$

$$Det(H) = D_{xx}D_{yy} - (D_{xy}^2) \tag{2.7}$$

$$\frac{Tr(H)^2}{Det(H)} < \frac{(r+1)^2}{r} \tag{2.8}$$

onde $Tr(H)$ corresponde ao traço da matriz (Equação 2.6), $Det(H)$ é o determinante da matriz Hessiana (Equação 2.7) e $r = 10$ por padrão do SIFT (valor encontrado, através de experimentos feitos pelo autor). Através do determinante da matriz Hessiana é possível saber o quanto a função está variando naquela região. Assim, são eliminados pontos de interesse que não são confiáveis, apesar de estarem em extremidades. Com isso, é obtido um conjunto de pontos de interesse com a localização espacial (x,y) e a escala σ correspondente.

Atribuição de uma orientação para cada ponto

Em seguida, é atribuída uma orientação para cada ponto de interesse com base na direção do gradiente local, a fim de tornar o descritor invariante a rotação. A orientação dos pontos é definida em três passos. O primeiro passo é determinar a magnitude e orientação do gradiente para cada pixel vizinho ao ponto de interesse. A magnitude calcula a força da variação no ponto e a orientação calcula a direção dessa variação. As Equações 2.9 e 2.10 descrevem as operações de magnitude m e orientação θ para um vizinho (x_σ, y_σ).

$$m(x_\sigma, y_\sigma) = \sqrt{(L(x_\sigma+1, y_\sigma, \sigma) - L(x_\sigma-1, y_\sigma, \sigma))^2 + (L(x_\sigma, y_\sigma+1, \sigma) - L(x_\sigma, y_\sigma-1, \sigma))^2} \tag{2.9}$$

$$\theta(x_\sigma, y_\sigma) = tan^{-1}\left(\frac{L(x_\sigma, y_\sigma+1, \sigma) - L(x_\sigma, y_\sigma-1, \sigma)}{L(x_\sigma+1, y_\sigma, \sigma) - L(x_\sigma-1, y_\sigma, \sigma)}\right) \tag{2.10}$$

onde σ representa a escala do ponto localizado. Calculada a magnitude e orientação dos pixels vizinhos, o segundo passo é calcular o histograma da orientação dos pixels vizinhos. O histograma é computado através do peso da magnitude de cada pixel vizinho. O propósito do histograma é identificar a orientação dominante na região ao redor do ponto de interesse. Lowe sugere que o histograma tenha 36 valores, para que os 360° de orientações possam ser representados. Por fim, extrai-se a orientação dominante a partir do pico do histograma computado. Caso haja uma orientação com pelo menos 80% do valor do pico, um novo ponto de interesse com a nova orientação é criado na mesma posição e escala do ponto de interesse original, pois essa também é uma orientação dominante na região (Lara, 2013).

Extração de descritores para cada ponto

Em seguida, é definida uma grade 4×4 rotacionada com a orientação calculada na etapa anterior em torno de cada ponto de interesse. Um histograma h_i das orientações dos pixels é calculado para cada grade i, como ilustrado na Figura 2.3. Foi determinado, experimentalmente por Lowe (2004), que cada ponto pode assumir 8 orientações diferentes, portanto, o histograma h_i gerado para cada grade tem 8 dimensões, isto é, $h_i \in \Re^8$. Os histogramas de cada grade são concatenados $h = [h_i] \ \forall \ i$, gerando um único descritor para cada ponto de interesse com 128 dimensões (4×4 regiões $\times$ 8 possíveis orientações). Por fim, o descritor de características é normalizado para reduzir as influências geradas por mudança de iluminação. O descritor é normalizado pela norma L2 (Equação 2.11). Em seguida, os valores do vetor são limitados em 0.2 se $h(i) > 0.2$ e, por fim, os descritores são normalizados novamente. O valor 0.2 foi encontrado experimentalmente pelos autores.

$$h = \frac{h}{||h||}, ||h|| = \sqrt{\sum_{i=1}^{128} |h(i)|^2} \tag{2.11}$$

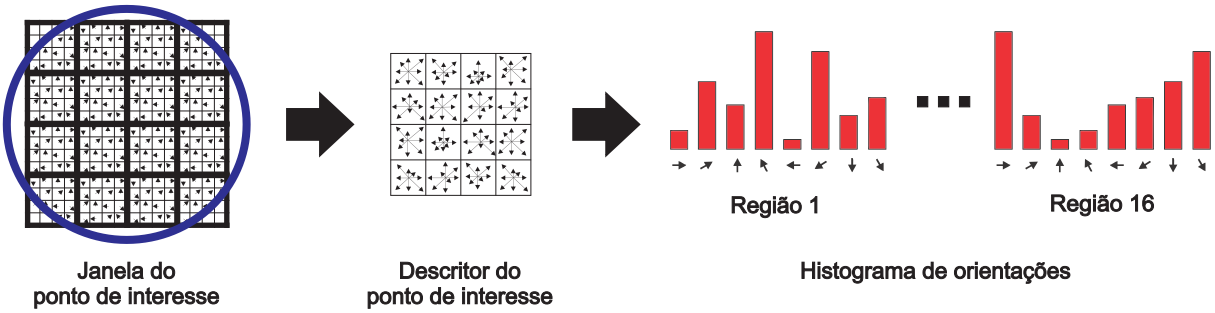


Figura 2.3: Uma grade de 4×4 é criada sobre o ponto de interesse, como mostrado à esquerda. Os pixels são ponderados por uma função Gaussiana, indicada pelo círculo sobreposto. Estes são, então, combinados em histogramas de orientação, como mostrado ao centro, para cada uma das 4×4 sub-regiões, onde o comprimento de cada seta correspondente à soma das magnitudes dos pontos localizados naquela região. A imagem mais a direita representa o histograma de orientações gerado por cada sub-região.

Ao final de todas as etapas, o SIFT extrai M pontos de interesse com 5 valores cada $\varphi_i = [x, y, \sigma, \theta, h]$, sendo (x, y) a posição espacial do ponto, σ a escala, θ a orientação e h são os descritores. A Figura 2.4 apresenta pontos de interesse encontrados em uma imagem através do descritor SIFT.

2.1.2 SIFT Denso

Após os bons resultados do SIFT (Se et al., 2002; Dorkó and Schmid, 2003), diversas variantes foram propostas, como PCA-SIFT (Ke and Sukthankar, 2004) e SIFT Denso (Liu et al., 2016). O SIFT Denso é um descritor que tem apresentado resultados promissores nas tarefas de reconhecimento de cenas e objetos (Liu et al., 2016). A principal diferença é que o SIFT Denso elimina a etapa de detecção de pontos e utiliza uma grade densa de pontos de interesse. Por exemplo, todos os pixels da imagem são considerados pontos de interesse ou uma grade que considera de b em b pixels, onde b é o parâmetro de densidade.

O SIFT em sua versão original, utiliza a etapa de detecção para encontrar pontos de interesse invariantes a escala e rotação. Desse modo, o SIFT Denso não é invariante a essas transformações, porém, são gerados mais descritores do que o SIFT original. A Figura 2.5 ilustra um exemplo de grade densa em uma imagem onde cada círculo representa um ponto de interesse. Assim, cada ponto desta grade é descrito da mesma maneira que o SIFT por

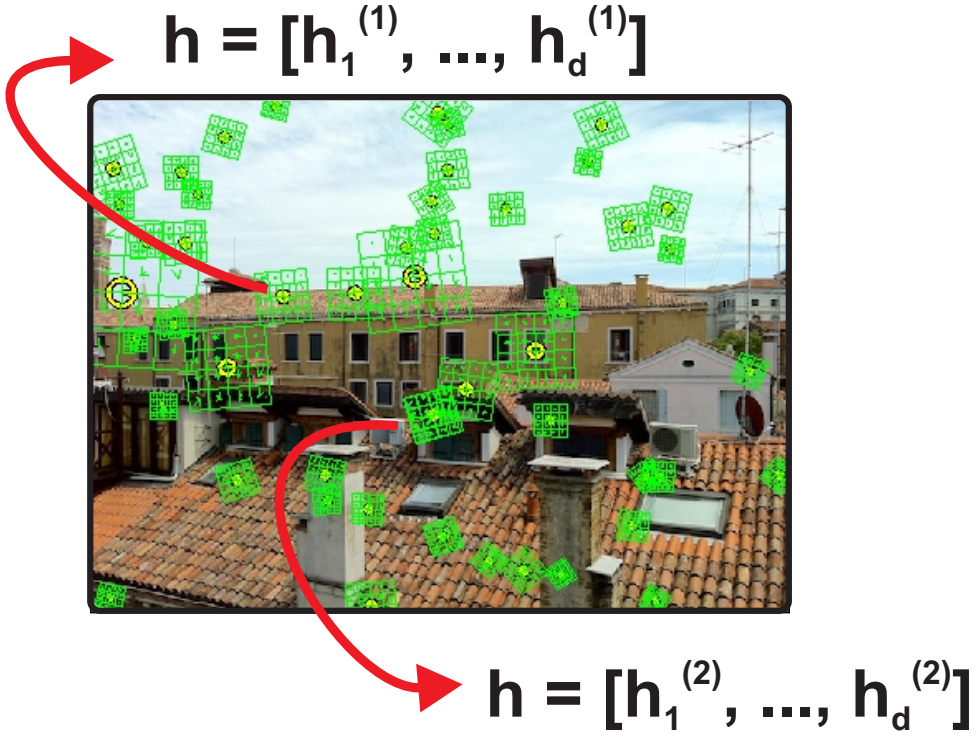


Figura 2.4: Exemplo do processo de detecção e descrição de pontos de interesse em uma imagem, feitos pelo algoritmo SIFT. Na última etapa, é extraído um vetor de características para cada ponto de interesse detectado na imagem.

meio do gradiente da região vizinha.

Os descritores locais do SIFT Denso são extraídos de forma semelhante ao SIFT original, porém levando em conta que todos os pixels da imagem são pontos de interesse. Primeiramente, é definido uma região de 16 $\times$ 16 pixels em torno do ponto de interesse. Então é calculada a magnitude e orientação para cada pixel vizinho, Equações 2.12 e 2.13, respectivamente. Em seguida, a região é dividida em uma grade 4 $\times$ 4, e posteriormente é calculado o histograma de 8 direções para cada grade. Por fim, os histogramas são concatenados, formando assim o descritor local do ponto de interesse. Assim como o SIFT original, o descritor extraído pelo SIFT Denso para cada ponto de interesse tem 128 dimensões.

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \tag{2.12}$$

$$\theta(x,y) = tan^{-1}\left(\frac{L(x,y+1) - L(x,y-1)}{L(x+1,y) - L(x-1,y)}\right) \tag{2.13}$$

A grande vantagem do SIFT Denso é que os pontos de interesse cobrem toda a imagem. Desse modo, o SIFT Denso gera descritores para todos os pixels da imagem, implicando um ganho de informações para caracterização, em comparação com SIFT tradicional que detecta regiões da imagem para serem descritas. Com isso, o número de pontos depende exclusivamente da grade utilizada na imagem. Por exemplo, se todos os pixels forem considerados como pontos de interesse, para uma imagem com $d \times a$ pixels são obtidos $M = d \times a$ pontos de interesse, onde d é a largura e a a altura da imagem. Desse modo, cada ponto de interesse é representado por $\varphi_i = [x, y, \theta, h]$, onde (x, y) é a localização espacial do ponto, θ é a orientação e h são os descritores locais.

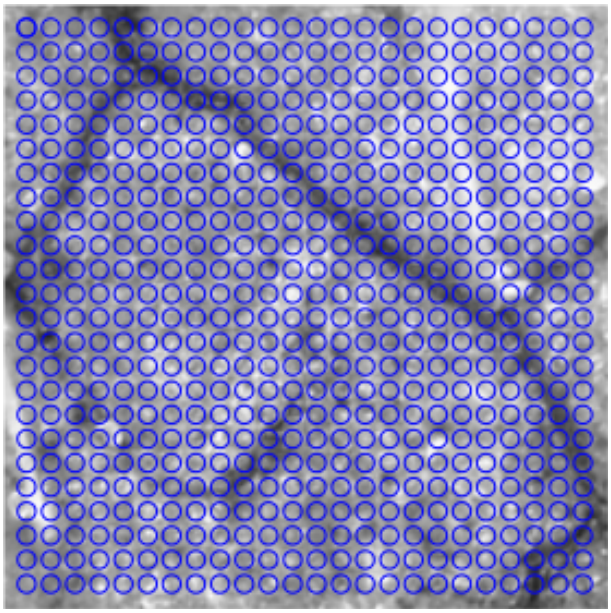


Figura 2.5: Exemplo de grade utilizada no SIFT Denso. Cada círculo representa um ponto de interesse na grade densa.

2.1.3 Histograma Piramidal de Palavras Visuais - PHOW

Apesar dos resultados promissores do SIFT Denso, essa abordagem não extrai descritores invariantes a escala. Dessa forma, é possível que duas imagens de um mesmo objeto, em diferentes escalas sejam classificadas como objetos diferentes. Com isso, foi proposto por Bosch et al. (2007) o Histograma Piramidal de Palavras Visuais (do inglês, *Pyramidal Histogram of Visual Words − PHOW*). O PHOW é uma abordagem que introduz a invariância à escala ao SIFT Denso (Seção 2.1.2), extraindo a magnitude e orientação em 4 escalas diferentes da imagem, conforme as Equações 2.14 e 2.15, respectivamente.

$$m(x,y,\sigma) = \sqrt{(L(x+1,y,\sigma) - L(x-1,y,\sigma))^2 + (L(x,y+1,\sigma) - L(x,y-1,\sigma))^2} \tag{2.14}$$

$$\theta(x,y,\sigma) = tan^{-1}\left(\frac{L(x,y+1,\sigma) - L(x,y-1,\sigma)}{L(x+1,y,\sigma) - L(x-1,y,\sigma)}\right) \tag{2.15}$$

Com isso, são gerados 4 descritores para cada ponto interesse em diferentes escalas σ. Os descritores são extraídos da mesma forma que o SIFT original, ou seja, com base em sua vizinhança. Para cada ponto de interesse, em uma determinada escala σ, é criada uma região 16×16, onde é calculada a magnitude e orientação de cada ponto. Por meio dessa região, são definidas grades 4×4, onde é computado o histograma de orientações para cada uma das regiões da grade. Por fim, os histogramas são concatenados e, como cada histograma tem 8 direções, temos um descritor local de tamanho 128.

O PHOW tem como principal vantagem, incluir informações invariantes a escala ao descritor SIFT Denso. O descritor PHOW, é formado por 4 descritores locais dos pontos de interesse da imagem obtidos em escalas diferentes. Desse modo, cada ponto de interesse é representado por $\varphi_i = [x,y,\theta,\sigma,h]$, onde (x,y) é a localização espacial do ponto, θ é a orientação, σ é a escala e h são os 4 descritores locais.

Uma variante ao PHOW foi proposta por Bosch et al. (2007), conhecida como PHOW-RGB. O PHOW-RGB consiste em aplicar o PHOW nos 3 canais de cor RGB, e assim, incluir informações de cor ao descritor. O descritor local extraído pelo PHOW-RGB, também é formado por 4 descritores locais dos pontos de interesse da imagem obtidos em escalas diferentes, porém,

dado um ponto de interesse em uma determinada escala, é extraído um descritor local para cada canal de cor. Em seguida, eles são concatenados para formar o descritor local do ponto de interesse em uma determinada escala, sendo assim, os descritores extraídos pelo PHOW-RGB tem um tamanho de 384, pois é extraído um descritor local de tamanho 128 para cada canal de cor RGB.

2.2 Bag-of-Visual-Words - BOVW

O histograma de palavras visuais, do inglês *Bag-of-Visual-Words* – BOVW, é uma técnica proposta por Csurka et al. (2004), utilizada em reconhecimento de cenas e objetos. Essa técnica foi inspirada no método *Bag-of-Words* de categorização de textos proposto por Joachims (1998). O método *Bag-of-Words* parte da ideia que textos podem ser classificados a partir da frequência de palavras (Lara, 2013). Da mesma forma, o BOVW categoriza imagens através da frequência de palavras visuais.

Basicamente, as palavras visuais são definidas por regiões de grande importância. No caso das imagens, as regiões importantes são definidas pelos pontos de interesse detectados por algum descritor local. Para definir as palavras visuais, é necessária a construção do vocabulário de palavras visuais, que é feito através da identificação de pontos de interesse extraídos de um conjunto de imagens para treinamento. Com o vocabulário de palavras visuais, são extraídos os pontos de interesse a partir de uma nova imagem e, cada ponto é rotulado em alguma palavra visual do vocabulário por meio de sua proximidade com a mesma. Desse modo, é possível computar a frequência de palavras visuais em uma imagem, a fim de reconhecê-la.

O BOVW é segmentado em quatro etapas principais: extração e descrição de pontos de interesse na imagem, geração do vocabulário visual, rotulação de cada ponto de interesse em uma palavra visual e combinação das palavras visuais da imagem em um histograma. As etapas do BOVW, são detalhadas a seguir.

2.2.1 Extração e descrição de pontos de interesse da imagem

A primeira etapa do BOVW tem como objetivo detectar e descrever pontos de interesse em uma imagem. Os pontos de interesse são definidos através das regiões consideradas importantes em uma imagem, e então, é extraído um vetor de características para cada um deles. Eles devem ser preferencialmente invariantes a transformações geométricas (e.g., escala, rotação) e transformações de iluminação, para que essas mudanças não interfiram nos descritores extraídos, e desse modo, a categorização não ser feita de forma incorreta.

Existem vários métodos que podem ser utilizados nessa etapa (e.g., SURF (Bay et al., 2006), SIFT (Lowe, 2004)). Apesar do SURF (Características Robustas Aceleradas, do inglês - *Speeded Up Robust Features*) ter apresentado ótimos resultados, um desempenho computacional melhor que o SIFT pelo fato de processar um descritor de características com uma dimensão menor (64) (Bay et al., 2006), foi mostrado por Mikolajczyk and Schmid (2005), que o SIFT tem uma melhor taxa de classificação correta, sendo o mais utilizado com o BOVW (Bolovinou et al., 2013; Zagoris et al., 2014).

Com a detecção e descrição dos pontos de interesse, obtemos um conjunto H de descritores locais que representa uma imagem I, como visto na Equação 2.16.

$$H_I = [h_1, h_2, ..., h_M]^T \in R^{M \times S}, \tag{2.16}$$

onde M representa a quantidade de pontos de interesse encontrados na imagem e S representa a dimensão dos descritores de cada ponto. É importante destacar que dependendo do

descritor utilizado, a quantidade de pontos de interesse M varia. Além disso, um descritor que utiliza a etapa de detecção em seu processo, pode detectar quantidades de pontos de interesse distintas em imagens que possuem o mesmo tamanho.

2.2.2 Geração do vocabulário visual

Após obter os descritores das imagens de treinamento, os mesmos são concatenados em um conjunto H. Esse conjunto é formado pelos descritores locais de todas as imagens de treinamento (Equação 2.17), como visto na Figura 2.6a.

$$H = [H_1, H_2, ..., H_n], \tag{2.17}$$

onde n representa a quantidade de imagens de treinamento. Um método de agrupamento é utilizado para agrupar os descritores locais em grupos. Geralmente o método usado é o *k-means* (Wu et al., 2008), por ser simples, intuitivo, produzir resultados fáceis de serem interpretados, além de ter complexidade linear. O *k-means* então é aplicado no conjunto de descritores locais H para obter um conjunto de centróides C (Figura 2.6b):

$$C = \text{k-means}(H), \tag{2.18}$$

onde k é o número de centróides. O centróide de cada grupo corresponde a média dos valores de todos descritores pertencentes aquele grupo. No *k-means*, a principio os k centróides são definidos aleatoriamente, em seguida cada descritor do conjunto de entrada é associado ao centróide mais próximo e por fim cada centróide tem seu valor atualizado pela média de todos os descritores associados ao seu grupo. Esse processo é feito até que o centróide tenha seu valor estabilizado ou o limite de iterações seja extrapolado.

Cada centróide corresponde a uma palavra visual e o conjunto das palavras visuais formam o vocabulário visual. Dessa forma, é possível quantificar a frequência de palavras visuais presentes em uma determinada imagem. As Figuras 2.6b e 2.6c ilustram o resultado do agrupamento feito pelo *k-means* e o conjunto de palavras visuais (vocabulário), respectivamente.

2.2.3 Rotulação de cada ponto de interesse em uma palavra visual

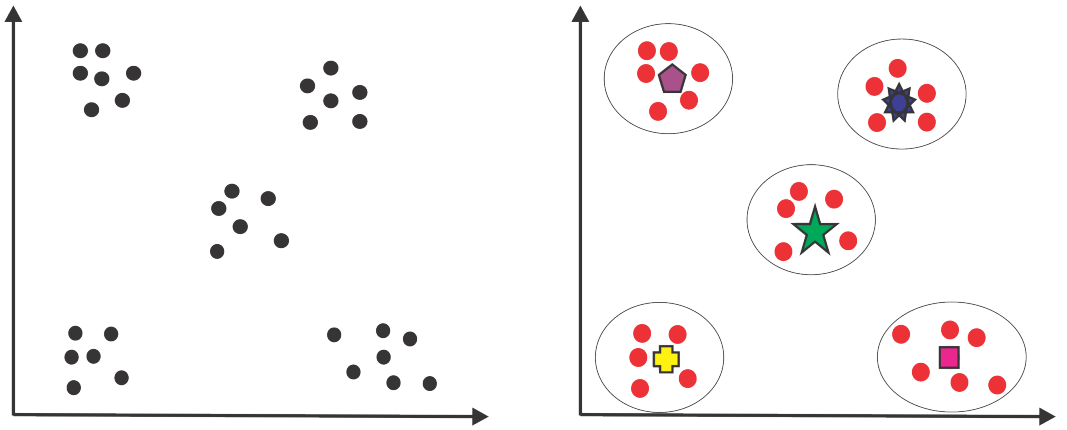
Para uma imagem ser representada por um vocabulário visual, cada ponto de interesse extraído da imagem é rotulado em uma palavra do vocabulário visual. O ponto de interesse é rotulado pela palavra visual que apresentar a menor distância Euclidiana entre seu descritor local h_i e o centróide de cada palavra do vocabulário visual:

$$R_i = arg \min_{j=0}^{k}(dist(h_i, C_j)) \tag{2.19}$$

onde $dist(.)$ é a distância Euclidiana entre o descritor h_i e a palavra visual C_j. A Figura 2.7 apresenta um exemplo de imagem com seus pontos de interesse rotulados em palavras do vocabulário visual.

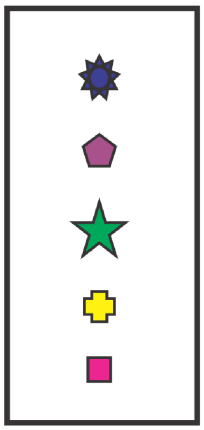
2.2.4 Combinar as palavras visuais da imagem em um histograma

Após cada descritor local da imagem ser rotulado em uma palavra visual, elas são combinadas em um histograma da frequência de palavras visuais, formando assim o descritor final da imagem, como mostrado na Figura 2.8. A dimensão do histograma de palavras visuais depende exclusivamente de k, pois este representa a quantidade de palavras do vocabulário visual.



(a) Conjunto de descritores locais das imagens de treinamento em um espaço S-dimensional.

(b) Exemplo da construção do vocabulário de palavras visuais. Os descritores locais das imagens de treinamento são agrupados em torno de k centróides.



(c) O vocabulário de palavras visuais é formado pelos k centróides.

Figura 2.6: Ilustração das etapas de agrupamento e construção do vocabulário de palavras visuais do método BOVW. Os descritores são extraídos das imagens de treinamento e concatenados em um espaço S-dimensional. Então, os descritores são agrupados em k grupos e os centróides dos k grupos formam o vocabulário de palavras visuais. No exemplo, foi utilizado um vocabulário de palavras visuais com $k = 5$.

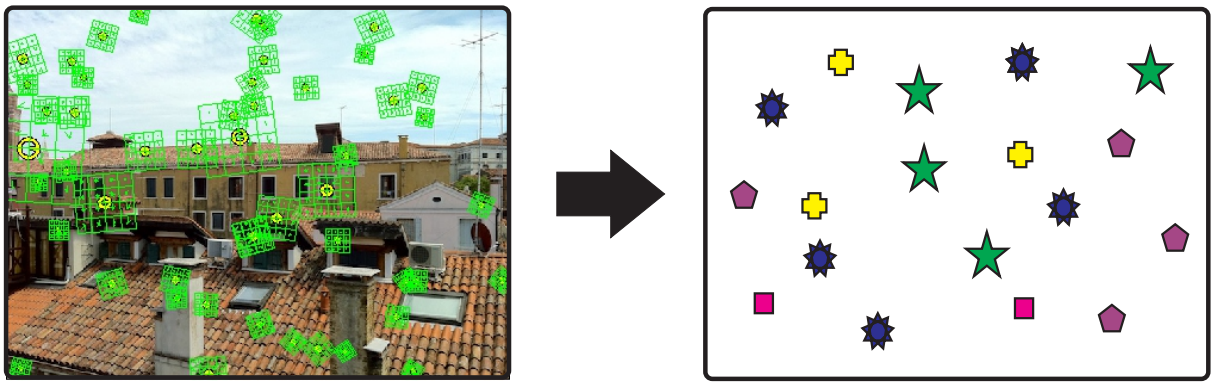


Figura 2.7: Exemplo de pontos de interesse identificados e rotulados em palavras do vocabulário visual. No exemplo, foi utilizado um vocabulário de palavras visuais de tamanho $k = 5$.

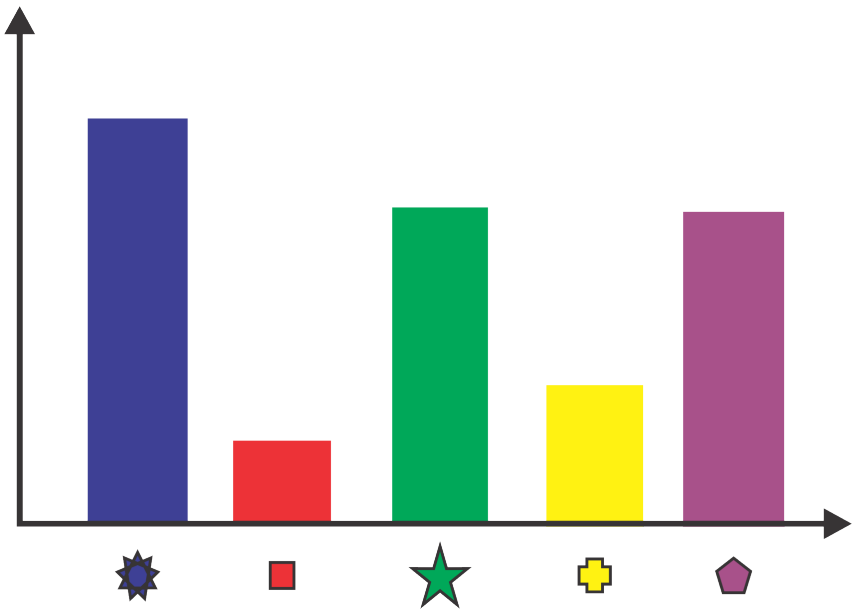


Figura 2.8: Exemplo do histograma da frequência de palavras visuais que representa o descritor final da imagem. O histograma foi computado através da imagem de exemplo apresentada na Figura 2.7. Nesse exemplo, foi utilizado um vocabulário de palavras visuais de tamanho $k = 5$.

2.3 Pirâmide Espacial

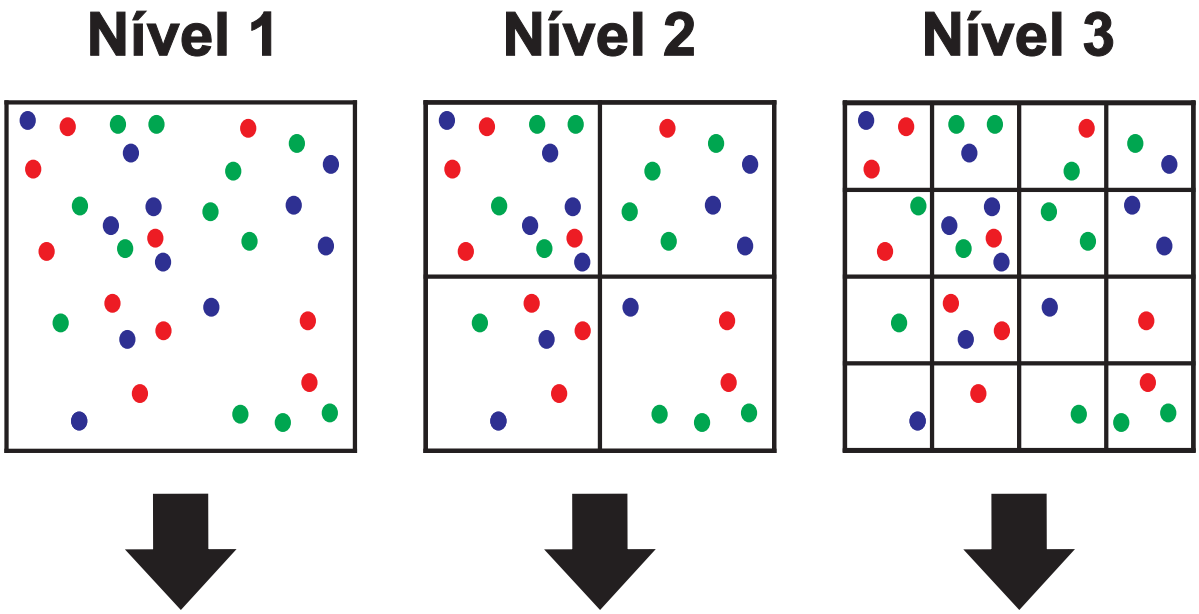
O método *Bag-of-Visual-Words* não inclui informações espaciais em seu descritor final, gerando incentivo a pesquisas para analisar a importância da informação espacial no reconhecimento de cenas e objetos. Um dos principais trabalhos nesta linha foi proposto por Lazebnik et al. (2006), onde o descritor é formado através de pirâmide espacial. Esse método divide a imagem em sub-regiões e calcula o histograma de palavras visuais de cada sub-região. A concatenação do histograma de cada sub-região gera o descritor final.

O método aplica a divisão da imagem duplicando o número de divisões a cada nível. No primeiro nível da pirâmide, $l = 0$, a imagem permanece em seu formato original, isto é, sem divisões conforme a Figura 2.9a. Portanto, no nível 0, o descritor da pirâmide espacial é igual ao descritor extraído pelo BOVW. Para o nível 1, a imagem é divida em 4 regiões, por meio da divisão dos eixos x e y em dois, conforme ilustrado na Figura 2.9b. O nível 2 divide cada uma das regiões do nível 1 em 4 outras regiões, totalizando 16 regiões (Figura 2.9c). Mais precisamente, a imagem é dividida em 4^l resoluções, onde l é o parâmetro correspondente ao nível da pirâmide. Para obter a caracterização de diferentes regiões da imagem, a mesma é dividida em sub-regiões seguindo uma pirâmide espacial. Para cada sub-região r_i o histograma $h^l_{r_i}$ é calculado. Em seguida, eles são concatenados conforme a Equação 2.20. A Figura 2.9 ilustra como os histogramas são concatenados.

$$H^l = [h_{r_0}, h_{r_1}, ..., h_{r_{4^l}}] \tag{2.20}$$

onde l representa o nível da pirâmide espacial e h_{r_i} é o histograma de todos pontos de interesse na região r_i. Através de experimentos dos autores, foi definido que o nível máximo é $l = 2$ e o tamanho do vocabulário de palavras visuais $k = 400$ utilizado no processo do BOVW.

Além dos histogramas concatenados em um determinado nível l de pirâmide, os histogramas dos níveis anteriores são incluídos em seu descritor final, como mostrado na Equação 2.21. Desse modo, o tamanho do descritor final da imagem será $(4^l k + 4^{l-1} k + ... + 4^0 k)$, ou seja,



(a) $H^0 = [h_{r_1}]$ (b) $H^1 = [h_{r_1}, h_{r_2}, ..., h^{r_4}]$ (c) $H^2 = [h_{r_1}, h_{r_2}, ..., h_{r_{16}}]$

Figura 2.9: Exemplos de vetores de características extraídos a partir da pirâmide espacial com $l = 0, 1, 2$.

muito maior que o tamanho do descritor gerado pelo BOVW que tem tamanho k.

$$H = [H^l, H^{l-1}, ..., H^0]. \tag{2.21}$$

2.4 Grafos

Apesar dos resultados promissores da pirâmide espacial, este método possui alguns problemas, como não ser invariante a rotação. Para contornar tais problemas, este trabalho propõe um método para inclusão de informações espaciais usando grafos. Portanto, esta seção descreve os fundamentos de grafos necessários para compreensão do trabalho desenvolvido. Um grafo é uma forma abstrata de representar soluções para vários problemas. O grafo é definido por um par $G = (V, E)$, onde $V = \{1, 2, ..., n\}$ é um conjunto finito de vértices e $E = \{e(i, j)\}$ é um conjunto de arestas que conecta um par de elementos não ordenados, não necessariamente distintos, de V (Goldbarg, 2012). Se $e(i, j)$ é uma aresta no grafo, dizemos que $e(i, j)$ está conectando i em j.

O grau de um vértice $d(i)$ corresponde ao número de arestas incidindo nele (Equação 2.22). No grafo apresentado na Figura 2.10a, o grau de todos os vértices são $d(i) = 2$. Sendo assim, um grafo regular é um grafo onde todos os vértices tem o mesmo grau $d(i)$ e não existe direcionamento entre os vértices incidentes, como podemos ver na Figura 2.10a. Um grafo direcionado são aqueles onde as arestas tem direcionamento, ou seja, se existe uma aresta $e(i, j)$ é possível ir de i para j, mas não é possível o caminho contrário pela mesma aresta. A Figura 2.10b, apresenta um exemplo de grafo direcionado, em suas representações toda aresta tem uma seta indicando o direcionamento. A ordem de um grafo corresponde ao número de vértices $|V(G)|$ existentes no grafo.

$$d(i) = \sum_{j=1}^{n} \begin{cases} 1, & \text{se } e(i, j) \text{ existe,} \\ 0, & \text{caso contrário,} \end{cases} \tag{2.22}$$

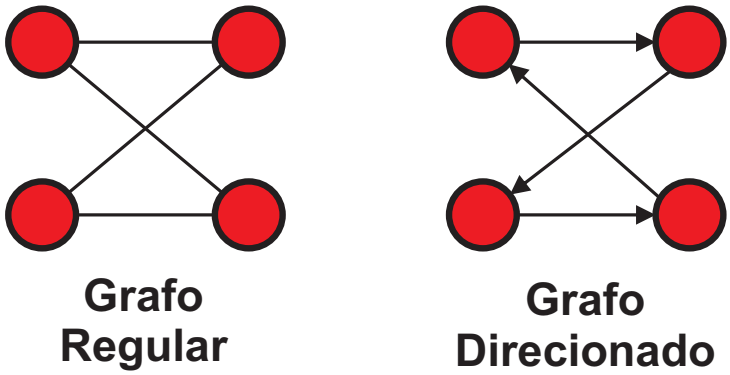


Figura 2.10: Exemplos de grafo não-direcionado (regular) e grafo direcionado.

Os grafos são computacionalmente representados por matrizes de adjacência, como mostrado na Figura 2.11. A Equação 2.23 apresenta como a matriz de adjacência deve ser completada. Na matriz de adjacência, quando há uma conexão entre dois vértices i e j, ou seja, existe uma aresta $e(i,j)$, então $A[i,j] = 1$, caso contrário, a matriz é preenchida com $A[i,j] = 0$.

$$A[i,j] = \begin{cases} 1, & \text{se} \quad e(i,j) \text{ existe,} \\ 0, & \text{caso contrário.} \end{cases} \tag{2.23}$$

Grafos ponderados são usados para representar conexões entre pares de vértices com pesos nas arestas. Desse modo, as arestas podem ser representadas utilizando valores reais. Por exemplo, se um grafo representa rodovias que conectam cidades, o peso pode ser o tamanho de cada rodovia. A Equação 2.24 apresenta como a matriz de adjacência deve ser preenchida.

$$A[i,j] = \begin{cases} w_{i,j}, & \text{se} \quad e(i,j) \text{ existe,} \\ 0, & \text{caso contrário.} \end{cases} \tag{2.24}$$

onde $w_{i,j}$ representa o peso da aresta $e(i,j)$.

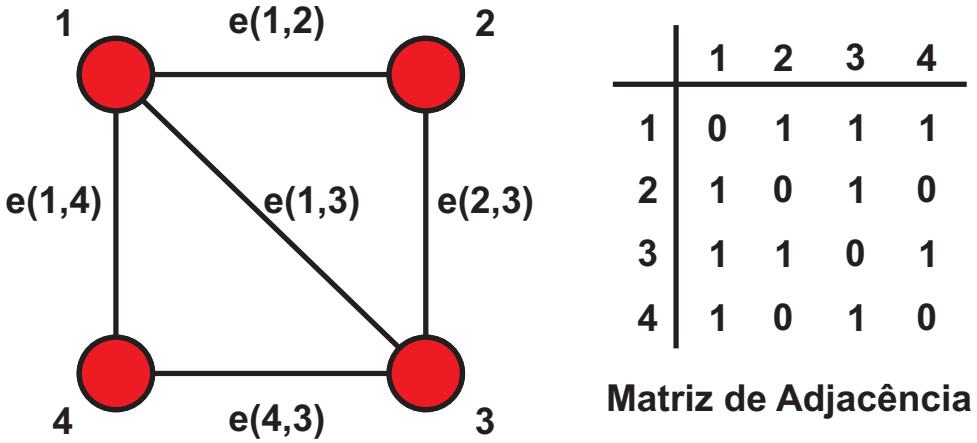


Figura 2.11: Representação computacional de um grafo por uma matriz de adjacência. Em uma matriz de adjacência quando dois vértices estão conectados (e.g., i e j), a posição $\{i,j\}$ na matriz de adjacência tem o valor 1 ou um valor correspondente ao peso dessa conexão.

CAPÍTULO

3

Metodologia

Neste capítulo, é apresentado o método proposto que tem como objetivo a inclusão de informações espaciais ao método BOVW, por meio de medidas extraídas de grafos. A Seção 3.1 apresenta a descrição do método proposto, onde primeiro é mostrado como a modelagem do grafo é executada e, em seguida é mostrado como as características são extraídas do grafo. Na Seção 3.2, são apresentadas as invariâncias à rotação e escala do método proposto. Posteriormente, a Seção 3.3 apresenta como o método de pirâmide espacial é incorporado ao método proposto. Por fim, na Seção 3.4 é apresentada a equivalência do método proposto com os métodos BOVW e pirâmide espacial.

3.1 Descrição do Método

Considere os N pontos de interesse extraídos pelo descritor local, em que cada ponto é representado por $p_i = (x_i, y_i, k_i)$, onde x_i e y_i representam a localização espacial do ponto p_i na imagem e k_i representa a palavra visual em que o ponto foi rotulado através da menor distância Euclidiana entre o descritor do ponto e as palavras visuais do vocabulário. A Seção 3.1.1 apresenta a modelagem desses pontos através de k grafos enquanto que a Seção 3.1.2 mostra como as características são extraídas para descrever a topologia do grafo.

3.1.1 Modelagem do Grafo

A proposta consiste em modelar um grafo G_w para cada palavra visual w, após a rotulação dos descritores locais em palavras visuais (Etapa 3 do BOVW). Para isso, os vértices de um grafo G_w correspondem aos pontos de interesse p_i rotulados na palavra visual w, isto é, $V_w = \{p_i | k_i = w\}$ representa o conjunto de vértices de uma determinada palavra visual w. Portando, o grafo $G_w = (V_w, E_w)$ modela somente pontos de interesse da palavra visual w. Para modelar todas as palavras visuais, são obtidos k grafos, um para cada uma das k palavras visuais.

A Figura 3.1 ilustra a forma como os grafos são modelados a partir das palavras visuais. No exemplo, temos a imagem na região superior representando a saída da terceira etapa do método BOVW, onde os pontos de interesse foram rotulados em palavras do vocabulário visual representadas por figuras geométricas. As imagens abaixo representam os grafos modelados

para cada palavra visual. Dessa forma, é possível extrair informações espaciais em relação a topologia de cada palavra visual na imagem. No exemplo, foi utilizado um vocabulário de palavras visuais com $k = 3$, desse modo, são modelados três grafos, um para cada palavra visual.

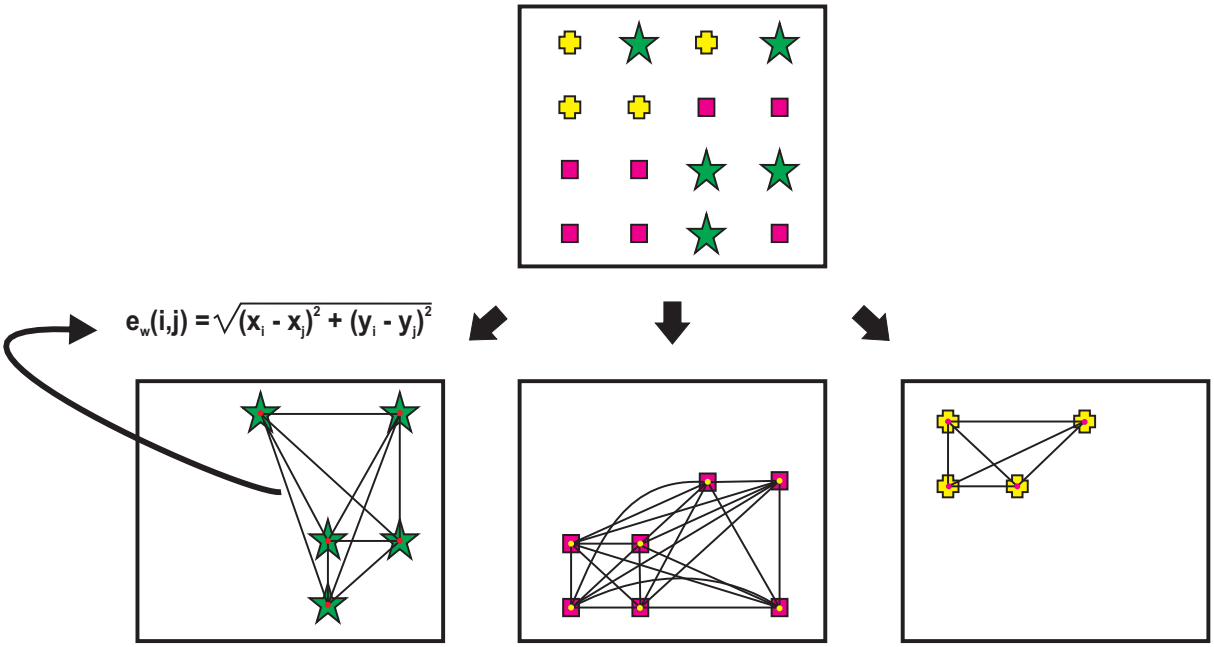


Figura 3.1: Exemplo da modelagem dos G_w grafos a partir das palavras visuais rotuladas. No exemplo, foi utilizado o vocabulário de palavras visuais com $k = 3$, sendo assim, são modelados 3 grafos, um para cada palavra visual w. Os pesos das arestas são representados pela distância Euclidiana entre os vértices.

Feito isso, é possível computar o peso da conexão entre um vértice i e j como sendo a distância Euclidiana entre a posição espacial dos pontos de interesse p_i e p_j: $e_w(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, onde (x_i, y_i) é a posição espacial do ponto de interesse i na imagem. Assim, o peso entre dois vértices i e j é grande se os pontos que eles representam, estão distantes espacialmente na imagem. Para tornar o grafo invariante ao tamanho da imagem e à escala, o peso das conexões é normalizado. Esse procedimento é formulado pela Equação 3.1 e tem o objetivo de tornar o método invariante a escala.

$$e_w(i,j) = \frac{e_w(i,j)}{max(E_w)} \tag{3.1}$$

onde $max(E_w)$ representa o valor máximo do conjunto de arestas E_w. Desse modo, as arestas $e_w(i,j)$ são normalizadas com valores entre $[0,1]$, sendo que $e_w(i,j) = 1$ se a distância entre p_i e p_j é máxima na imagem.

Para extrair diferentes características do grafo, é possível eliminar arestas do conjunto E_w conforme a Equação 3.2, onde um vértice é conectado à outro caso eles estejam a um raio r, isto é, $e_w(i,j) \leq r$.

$$e_w(i,j) = \begin{cases} e_w(i,j), & \text{se} \quad e_w(i,j) \leq r \\ -1, & \text{caso contrário.} \end{cases} \tag{3.2}$$

Esse corte de arestas tem como objetivo conectar vértices que estão próximos espacialmente na imagem, podendo assim, modelar o grafo em diferentes topologias e averiguar qual topologia resulta em uma melhor descrição da imagem. A Figura 3.2, ilustra um corte de arestas feito por um determinado raio r. No exemplo, foram utilizados os grafos da Figura 3.1, onde foram mantidas somente as arestas que conectam vértices vizinhos laterais, superiores

e inferiores. Como as arestas dos grafos são normalizadas entre $[0,1]$, o valor do raio r também está no intervalo $[0,1]$.

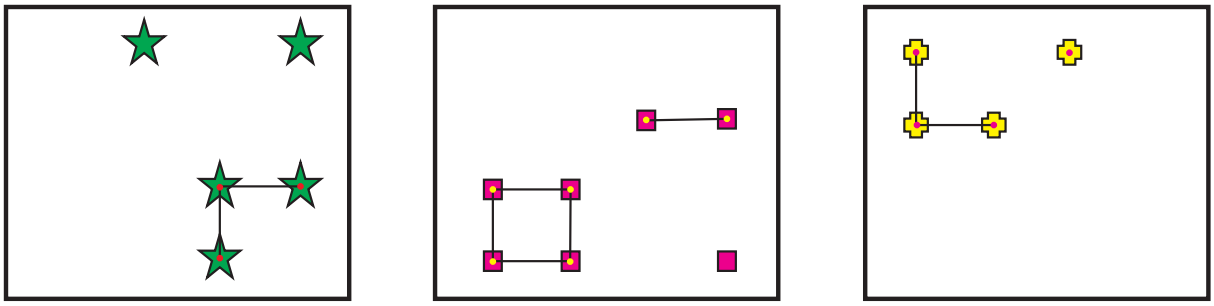


Figura 3.2: Exemplo de como é feito o corte de arestas que tenham um peso de conexão maior que um determinado raio r. Neste exemplo, foram utilizados os grafos da Figura 3.1 para fazer o corte de arestas. Nesse exemplo, foram mantidos apenas as arestas que conectam vértices vizinhos laterais, superiores e inferiores.

A Figura 3.3b ilustra um grafo modelado com $k = 2$ (círculos e quadrados) e raio $r = 0.8$, o grafo foi obtido a partir da imagem ilustrada na Figura 3.3a. Dado o grafo completo, é possível aplicar uma função para caracterizar diferentes propriedades do objeto. Esta função remove arestas cujo a distância Euclidiana é maior que o limiar r (Equação 3.2). As Figuras 3.3c e 3.3d ilustram a função de remoção de arestas aplicada no grafo mostrado na Figura 3.3b, utilizando raio $r = 0.6$ e 0.2, respectivamente. Para cada raio r, o grafo é transformado em um grafo de escala r que apresenta novas propriedades, como estrutura e topologia associadas à sua escala. Para valores pequenos de r, os detalhes da imagem são destacados por pequenos grupos de palavras visuais conectadas (Figura 3.3d). À medida que r aumenta, o grafo apresenta informações globais da imagem, como regiões onde o número de palavras visuais conectadas é maior.

3.1.2 Extração de Características do Grafo

A partir da modelagem dos grafos, são extraídas características em relação a topologia do grafo G_w, para formar o descritor final da imagem. Conforme explicado na Seção 3.1.1, um grafo é modelado para cada palavra visual w. Primeiramente, é calculado o grau $d_w(i)$ de cada vértice do grafo G_w, conforme a Equação 3.3. Os vértices que possuem grau alto, estão em regiões com mais vértices ao redor (região densa). Se levarmos em conta o método BOVW, isso representa que pontos de interesse com palavras visuais iguais estão próximos. Com isso, podemos extrair informações em relação a localização espacial dos vértices que pertencem a uma mesma palavra visual.

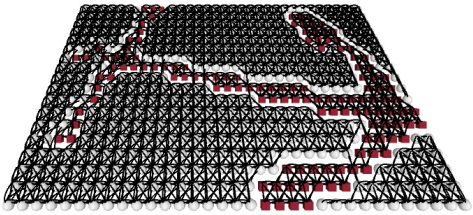
$$d_w(i) = \sum_{j=1}^{n} \begin{cases} 1, & \text{se} \quad e_w(i,j) \geq 0, \\ 0, & \text{caso contrário.} \end{cases} \tag{3.3}$$

Em seguida, é extraído o grau médio dos vértices $\mu(d_w) = \frac{1}{n}\sum_{i=1}^{n} d_w(i)$, e então, essa medida é utilizada como característica para o método proposto. Com o grau médio, temos a informação da proximidade de palavras visuais, em relação a posição espacial da mesma na imagem. Por último, o grau médio dos grafos de cada palavra visual são concatenados, formando assim, o descritor final F como mostrado na Equação 3.4.

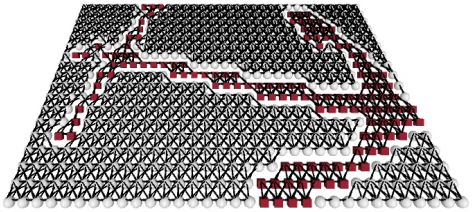
$$F = [\mu(d_1), \mu(d_2), ..., \mu(d_k)]. \tag{3.4}$$

Dessa forma, conforme a distribuição das palavras visuais pela imagem, somente são ex-

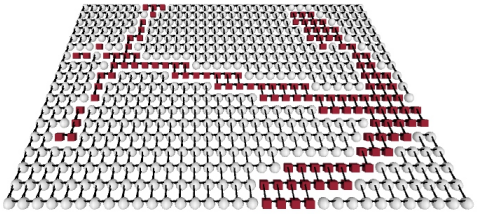
(a)



(b) $r = 0.8$



(c) $r = 0.6$



(d) $r = 0.2$

Figura 3.3: Exemplo da função de remoção de arestas aplicada em um grafo usando raio (b) $r = 0.8$, (c) $r = 0.6$, e (d) $r = 0.2$. No exemplo, foi utilizado o descritor local SIFT Denso para extrair os pontos de interesse da imagem.

traídas informações da topologia de grafos com palavras visuais iguais. A Figura 3.4 mostra como é o procedimento da extração de características do método proposto. A partir dos k grafos modelados, é calculado o grau dos vértices e em seguida, é extraído o grau médio de cada grafo. Por fim, o grau médio de todos os grafos são concatenados, formando assim, o descritor final da imagem. Portanto, temos o descritor final da imagem de tamanho k, onde k representa o número de palavras visuais utilizadas no vocabulário. Por meio do exemplo apresentado na Figura 3.4, podemos perceber que em grafos onde as palavras visuais estão próximas obtêm-se um grafo com mais conexões e consequentemente o grau médio tem um valor maior.

3.2 Invariâncias do Método

Nesta seção, são descritas as invariâncias do método proposto com relação à escala e rotação. Nos exemplos abaixo, considera-se o grafo para cada palavra visual sem perda de generalidade. Primeiro, é mostrado a invariância à rotação que a topologia de grafos apresenta. Posteriormente, é mostrado a invariância à escala, pelo fato da normalização das distâncias entre os vértices do grafo.

A partir da matriz de adjacência é possível extrair informações espaciais invariante à rotação. A Figura 3.5 apresenta a medida extraída pelo método proposto, onde dois grafos foram modelados a partir de imagens em diferentes rotações. Na Figura 3.5, são ilustrados dois grafos, ambos com 7 vértices e 9 arestas, e ao lado de cada vértice está o grau do mesmo, representado por $d(i)$. Com isso, é possível perceber que caso haja duas imagens em diferentes rotações, os vetores de características extraídos pelo método são os mesmos, mostrando a invariância à rotação do método. Entretanto, as características extraídas pelo método somente serão invariantes à rotação, caso o descritor local também seja, isto é, os pontos de interesse

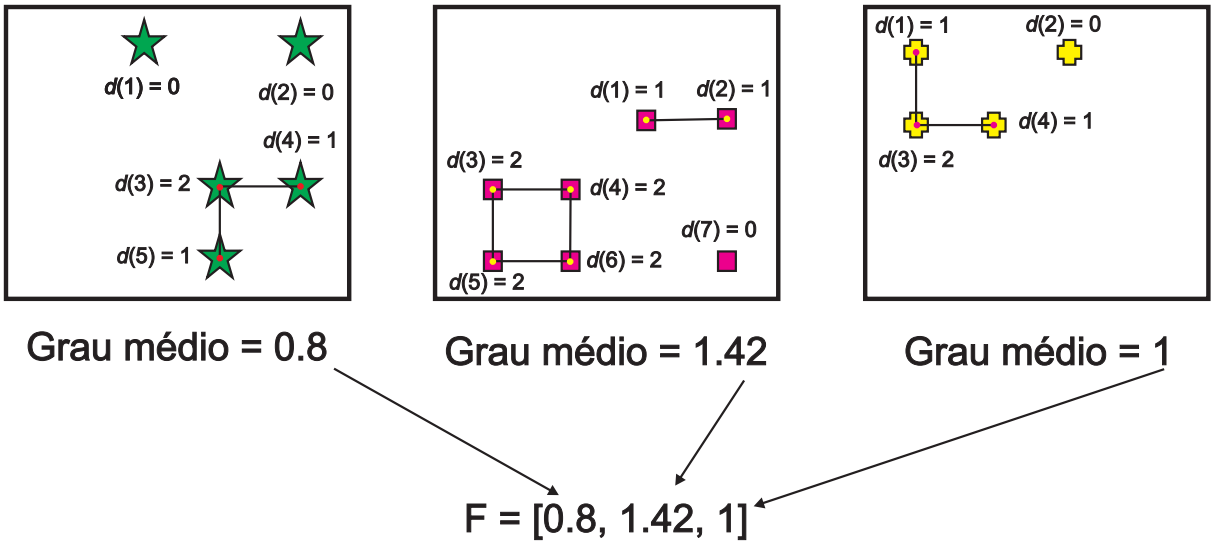


Figura 3.4: Exemplo da extração de características dos grafos. Primeiramente, é calculado o grau dos vértices de todos os grafos, na sequência, é computado o grau médio dos grafos que, por fim, são concatenados em um descritor.

detectados devem ser os mesmos. Conforme apresentado por Lowe (2004) e Liu et al. (2016), o SIFT, SIFT Denso e o PHOW apresentam essa importante característica.

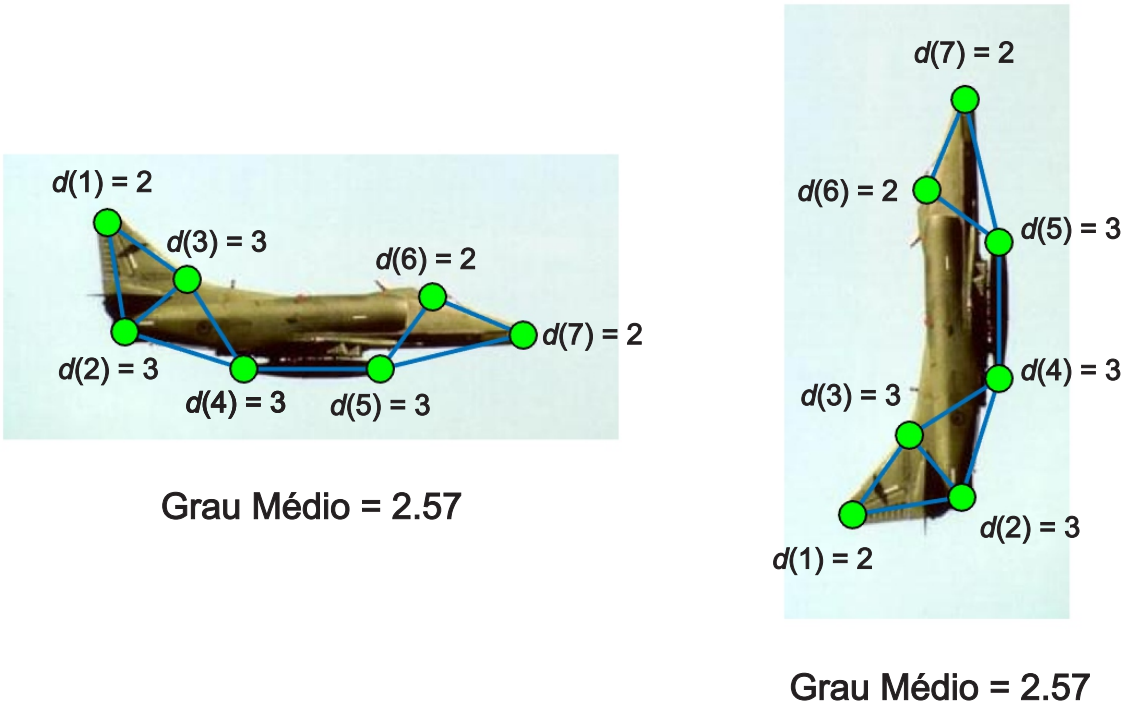


Figura 3.5: Exemplo de duas imagens em rotações diferentes, modeladas por grafos. Com o grau médio extraído pelo método proposto, é possível perceber a robustez do método em relação a rotação.

A invariância à escala é um constante problema no reconhecimento de cenas e objetos. No método proposto, a solução desse problema é realizada por meio da normalização do peso das arestas, ou seja, a distância entre os vértices. O peso das arestas é normalizado pela aresta de máxima distância no grafo. Dessa forma, considere duas imagens em diferentes escalas. Em princípio, os grafos modelados para as duas imagens terão arestas com pesos

diferentes, devido à distância espacial dos vértices. Com a normalização do peso das arestas pela aresta de maior distância no grafo, ambos os grafos terão as arestas com pesos iguais. A Figura 3.6 mostra um exemplo do funcionamento dessa normalização e como a medida extraída pelo método proposto é consistente, mesmo em imagens com escalas diferentes. Na Figura, temos como exemplo, a aresta de máxima distância (maior peso) e o peso de uma aresta selecionada. Desse modo, todas as arestas terão seus respectivos pesos normalizados pela aresta com maior distância no grafo, e assim todas as arestas terão como peso, valores entre $[0,1]$. Com isso, dadas duas imagens em escalas diferentes, o grafo modelado para imagens serão idênticos, pois a proporção das distâncias é mantida. Desse modo, o método proposto mostra ser robusto a escala.

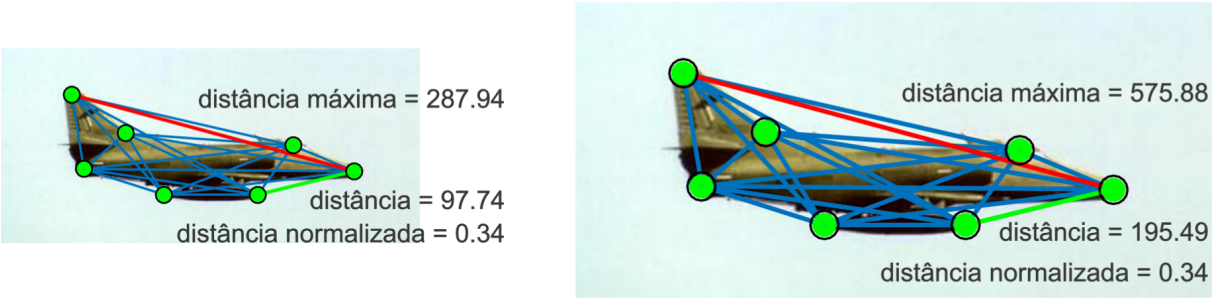


Figura 3.6: Exemplo de duas imagens em escalas diferentes. Nesse exemplo, as arestas são normalizadas pela aresta de máxima distância, ou seja, aresta com o maior peso. A linha vermelha representa a aresta com a maior distância no grafo e a linha verde representa a distância que está sendo normalizada.

3.3 Divisão em multi-resolução

A última etapa do método proposto consiste em dividir a imagem em multi-resoluções utilizando o algoritmo pirâmide espacial mostrado na Seção 2.3. Esta técnica divide a imagem em sub-regiões menores, calculando para cada sub-região um histograma de características locais. O método original aplica a divisão da imagem quadruplicando o número de divisões a cada nível. No primeiro nível, $l = 0$, a imagem permanece em seu formato original, isto é, sem divisões. Para o nível 1, a imagem é divida em 4 regiões, por meio da divisão dos eixos x e y em dois e assim por diante.

Em nosso método, ao invés de calcularmos o histograma para cada sub-região, é aplicado o método descrito na Seção 3.1. Desse modo, o método é aplicado em cada sub-região, ou seja, é criado um grafo $G^l_{w_i}$, para cada sub-região i do nível da pirâmide l. Sendo assim, podemos extrair informações da topologia em cada sub-região da imagem, fazendo com que seja explorada, ainda mais, as informações espaciais da imagem. Em nosso método, é extraído o grau médio de cada grafo, como mostrado na Seção 3.1.2. Desse modo, temos para cada sub-região um vetor F^l_i, que representa o vetor de características da sub-região i, com o nível de pirâmide l. Por fim, os vetores de cada nível da pirâmide são concatenados, formando o descritor final da imagem F. A Equação 3.5 mostra a formalidade do nosso método.

$$F^l = [F_i, F_{i+1}, ..., F_{4^l}], \qquad F = [F^1, F^2, ..., F^l]. \tag{3.5}$$

A Figura 3.7 apresenta um exemplo de como as características são extraídas utilizando nossa abordagem. Neste exemplo, foram utilizados como parâmetros, o vocabulário de palavras visuais $k = 3$ e o nível da pirâmide $l = 2$.

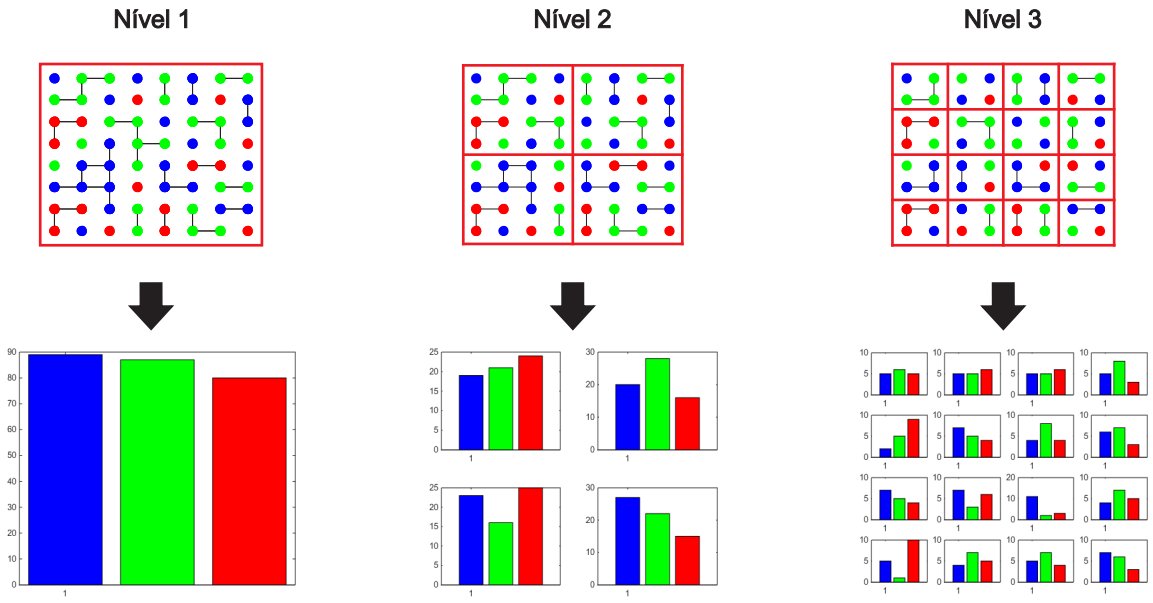


Figura 3.7: Exemplo da extração de características executada pelo método proposto. O grau médio $\mu(d_w)$ é calculado para cada sub-região em diferentes níveis da pirâmide. Por fim, os vetores obtidos por cada sub-região são concatenados, além de concatenar com os vetores obtidos das sub-regiões dos níveis anteriores.

3.4 Método proposto: uma generalização do BOVW e Pirâmide Espacial

O método proposto é uma generalização dos métodos BOVW e pirâmide espacial visto que, com ajustes de parâmetros do método proposto, é possível contemplar o BOVW e Pirâmide Espacial. Como exemplo, temos a Figura 3.8 que ilustra a aplicação do método BOVW. No exemplo, temos a esquerda uma imagem sintética, onde foi aplicada a rotulação dos descritores locais em palavras visuais, utilizando $k = 3$. Neste caso, as palavras visuais foram representadas por três figuras geométricas: estrela, quadrado e cruz. Aplicando o histograma na imagem, ou seja, computando a frequência de palavras visuais na imagem, temos 5 estrelas, 7 quadrados e 4 cruzes. Dessa forma, temos $H = [5,7,4]$ como o descritor final da imagem fornecido pelo método BOVW.

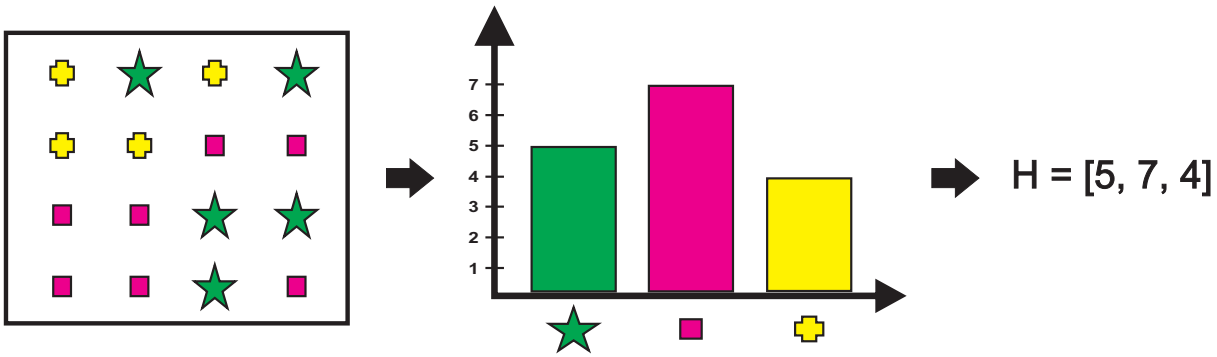


Figura 3.8: Exemplo da aplicação do método BOVW em uma imagem sintética. No exemplo, foi utilizado um vocabulário de palavras visuais de tamanho $k = 3$, onde as palavras visuais são representadas por figuras geométricas. No exemplo, foi considerado que todo vértice está conectado com ele mesmo, portanto, o grau de todo vértice foi incrementado em 1.

Agora, considere o exemplo mostrado na Figura 3.9, onde foi aplicado o método proposto. Primeiramente, foi feita a rotulação dos descritores locais em palavras visuais, utilizando um

vocabulário visual $k = 3$, em que cada palavra visual está representada por figuras geométricas. O método proposto foi aplicado com os parâmetros de raio $r = 1$ e nível da pirâmide $l = 0$, portanto, teremos como saída grafos completos, ou seja, todos os vértices estarão conectados com todos os outros, além de cada vértice estar conectado com ele mesmo. Por exemplo, no grafo que possui a estrela como a representação de uma palavra visual, existem 5 vértices, onde um vértice i está conectado com todos os outros e com ele mesmo, então $d_w(i) = 5$. Como todos os vértices tem o mesmo grau, o grau médio desse grafo é $\mu(d_w) = 5$, que corresponde a quantidade de palavras visuais de w, como visto na Figura 3.8. O nível da pirâmide $l = 0$ indica que não haverá divisões na imagem, como explicado na Seção 2.3.

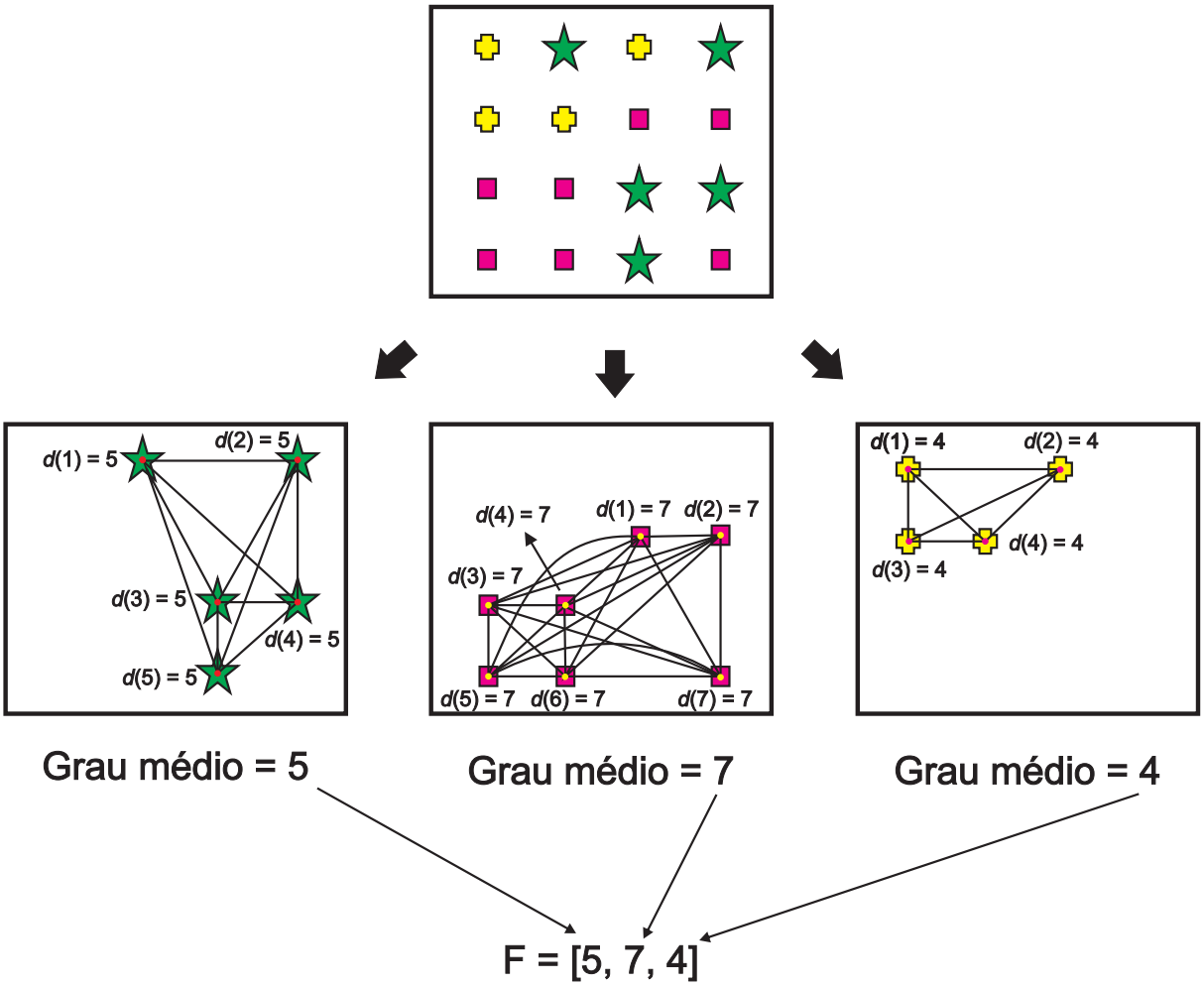


Figura 3.9: Exemplo da aplicação do método proposto em uma imagem sintética. No exemplo, foi utilizado um vocabulário de palavras visuais de tamanho $k = 3$, onde as palavras visuais são representadas por figuras geométricas. Os parâmetros do método nível da pirâmide e raio, foram setados com $l = 0$ e $r = 1$, respectivamente. No exemplo, foi considerado que todo vértice está conectado com ele mesmo, portanto, o grau de todo vértice foi incrementado em 1.

Desse modo, o descritor final da imagem extraído pelo método proposto é $F = [5,7,5]$. Com isso, o vetor de características do método proposto é igual ao vetor extraído pelo BOVW, caso o método proposto utilize os parâmetros raio $r = 1$ e nível da pirâmide $l = 0$, como mostrado nas Figuras 3.8 e 3.9.

Para mostrar a equivalência do método proposto e o pirâmide espacial, considere o exemplo da Figura 3.10 que ilustra o método de pirâmide espacial com o nível da pirâmide $l = 1$. O nível da pirâmide $l = 1$, indica que a imagem é dividida em 4 sub-regiões. No exemplo, a imagem com as palavras visuais rotuladas é a mesma que a utilizada nos exemplos das Figuras 3.8 e 3.9, onde foi utilizado um vocabulário de palavras visuais $k = 3$. Primeiramente, a imagem

foi dividida em 4 sub-regiões e, posteriormente, é aplicado o histograma de palavras visuais em cada uma dessas sub-regiões. Por fim, os histogramas das sub-regiões são concatenados, formando assim, o descritor final da imagem fornecido pelo método pirâmide espacial.

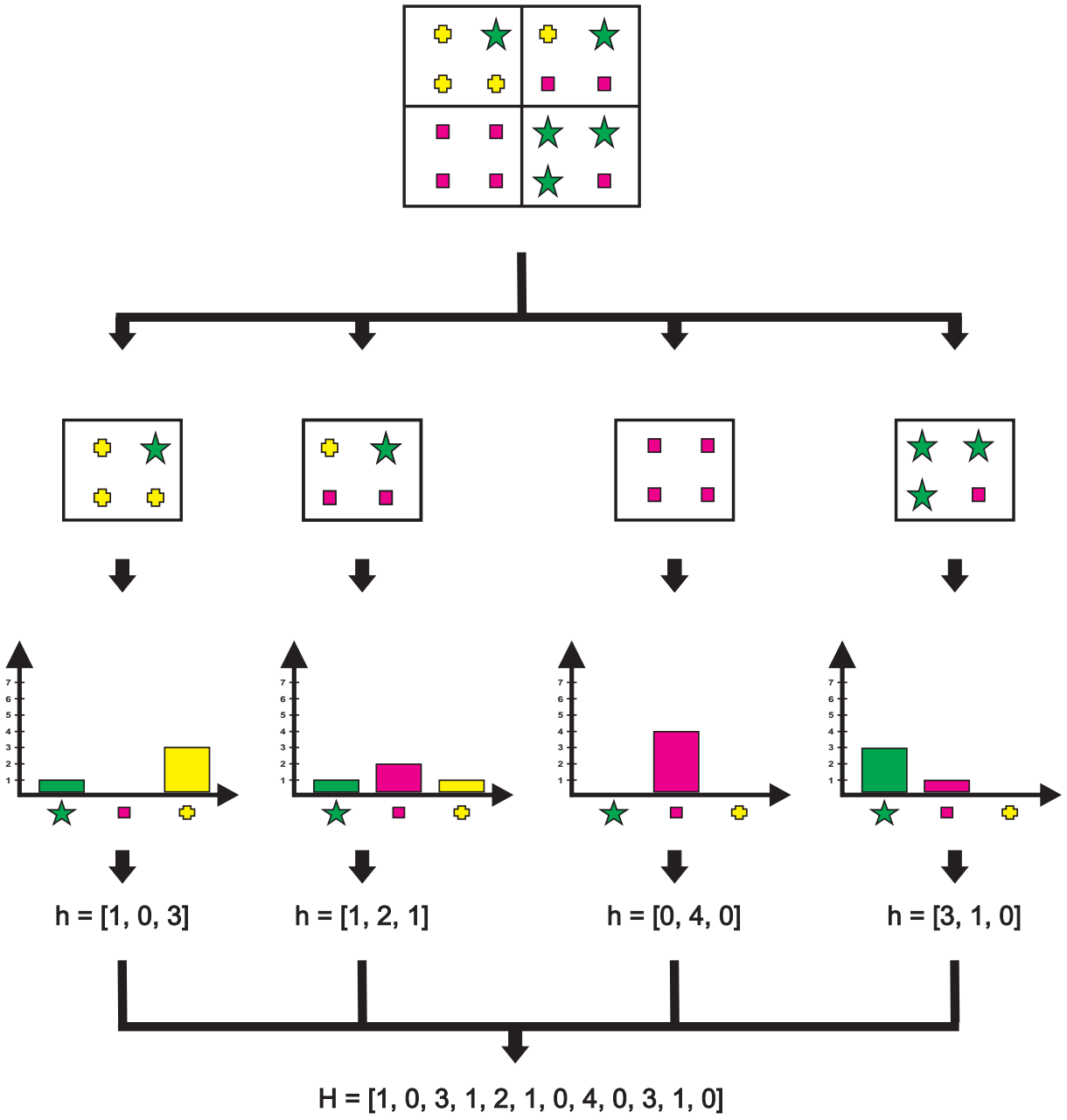


Figura 3.10: Exemplo da aplicação do método pirâmide espacial, utilizando o nível da pirâmide $l = 1$. Nesse exemplo, foi utilizada a mesma imagem sintética das Figuras 3.8 e 3.9, com $k = 3$. Primeiro, a imagem foi dividida em 4^l sub-regiões. Em seguida, foi calculado o histograma de palavras visuais h em cada uma dessas sub-regiões. Por fim, os histogramas h foram concatenados, formando o descritor final da imagem H.

Para mostrar a analogia do método proposto com o pirâmide espacial, a Figura 3.11 ilustra a aplicação do método proposto, onde a imagem com as palavras visuais rotuladas é mesma utilizada na Figura 3.10. Nesse exemplo, foram utilizados os parâmetros de raio $r = 1$ e nível da pirâmide $l = 1$. O raio $r = 1$, indica que não haverá cortes nas arestas, ou seja, os grafos modelados são completos. Com o nível da pirâmide $l = 1$, temos uma imagem dividida em 4 sub-regiões, assim como no método pirâmide espacial. Como o vocabulário de palavras visuais tem tamanho $k = 3$, são modelados 3 grafos, um para cada palavra visual. Os grafos modelados são completos, desse modo, todo vértice está conectado com todos os outros, além

de estar conectado com ele mesmo. Dessa forma, o grau de todos os vértices de um grafo G_w são iguais e equivalentes a quantidade de vértices que correspondem as palavras visuais daquele grafo. Posteriormente, é extraído o grau médio de cada grafo que, em seguida, são concatenados formando o descritor de uma sub-região. Por fim, os descritores de cada sub-região são concatenados formando o descritor final da imagem F. Como o método proposto extrai o mesmo número de características que o pirâmide espacial, o tamanho do descritor final da imagem será $(4^l k + 4^{l-1} k + ... + 4^0 k)$.

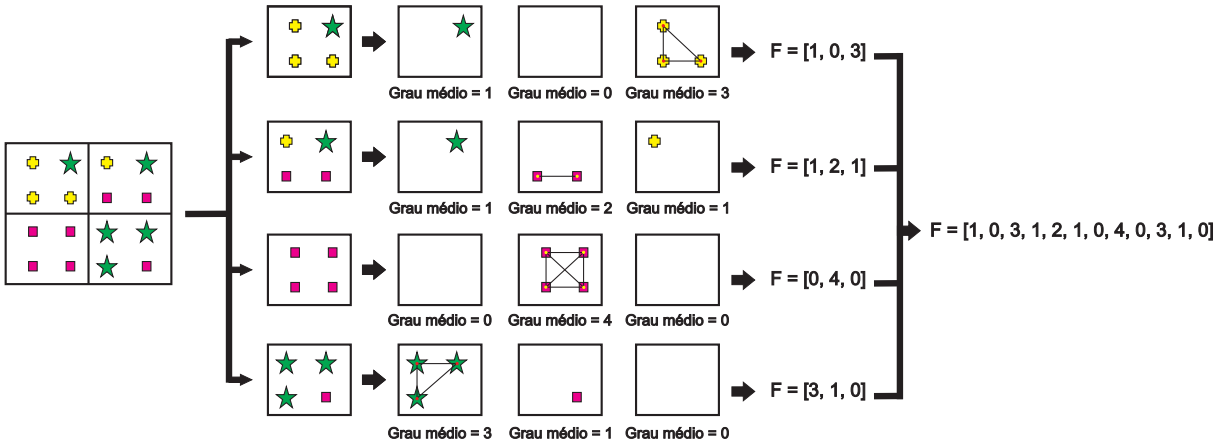


Figura 3.11: Exemplo da aplicação do método proposto, utilizando raio $r = 1$ e nível da pirâmide $l = 1$. Nesse exemplo, foi utilizada a mesma imagem sintética da Figura 3.10. Primeiro, a imagem foi dividida em 4^{l-1} sub-regiões. Em seguida, foram modelados 3 grafos em cada sub-região, um para cada palavra visual. Em seguida, é calculado o grau médio de cada grafo que, posteriormente são concatenados formando o descritor da sub-região. Por fim, os descritores de cada sub-região, são concatenados, formando assim, o descritor final da imagem F extraído pelo método proposto. No exemplo, foi considerado que todo vértice está conectado com ele mesmo, portanto, o grau de todo vértice foi incrementado em 1.

Através das Figuras 3.10 e 3.11, podemos perceber que para uma mesma imagem, utilizando os parâmetros raio $r = 1$ e nível da pirâmide $l = 1$, temos que o descritor final da imagem obtido pelo pirâmide espacial é equivalente ao descritor final da imagem extraído pelo método proposto. Com isso, o método proposto mostra ser equivalente aos métodos BOVW e pirâmide espacial. Contudo, variando o parâmetro raio r, teremos grafos com diferentes topologias, uma vez que, com baixos valores de r, teremos menos conexões e com altos valores de r teremos um grafo onde as palavras visuais distantes estarão conectadas, resultando em descritores diferentes dos métodos BOVW e pirâmide espacial.

De um modo geral, nossa técnica é formulada na generalização dos métodos BOVW e pirâmide espacial. Sendo assim, utilizando o nível da pirâmide $l = 0$ e raio $r = 1$, o grafo modelado (um para cada palavra visual) terá grau médio equivalente ao número de vértices do mesmo, pois o grafo será completo. Com isso, para cada grafo modelado, será extraído simplesmente a frequência com que cada palavra visual aparece na imagem, isto é, o mesmo conceito do BOVW. Desse modo, temos que nosso método extrai a mesma quantidade de características do pirâmide espacial, porém as características extraídas pelo método proposto, levam em consideração informações espaciais em cada sub-região.

CAPÍTULO

4

Experimentos e Resultados

Neste capítulo, são descritos os experimentos e os resultados obtidos. A Seção 4.1 apresenta os protocolos utilizados nos experimentos, como os descritores locais, o classificador, as medidas de avaliação e a base de imagens. A Seção 4.2 mostra a análise de parâmetros do método proposto, além de uma comparação entre o método proposto e os métodos BOVW e pirâmide espacial, o principal método da literatura que inclui informações espaciais ao BOVW.

4.1 Protocolo Experimental

Nos experimentos apresentados posteriormente, foi utilizada a base de imagens Caltech-101. As imagens dessa base, foram recolhidas em Setembro de 2003 por Fei-Fei et al. (2007). A base contém 9.144 imagens, separadas em 101 classes de objetos diferentes (e.g., faces, celulares, aviões, etc.) mais uma classe de fundo, totalizando 102 classes. A base possui cerca de 40 a 800 imagens por classe, embora a maioria das classes tenha em torno de 50 imagens. A Caltech-101 é um grande desafio, devido a elevada variabilidade dos elementos de uma mesma classe. A escolha dessa base, é pelo fato dos objetos estarem em diferentes rotações e escalas, como podemos perceber na Figura 4.1. A Figura 4.3 ilustra um exemplo de cada classe de imagens contidas nessa base.

a) âncora b) formiga c) castor d) lagosta

Figura 4.1: Exemplo de 4 classes (âncora, formiga, castor e lagosta) presentes na base Caltech-101. É possível perceber a variância entre as amostras de uma mesma classe, como a rotação e escala, intensificando a dificuldade no aprendizado dos métodos de reconhecimento.

As imagens tiveram seus descritores locais extraídos a partir dos métodos SIFT Denso, PHOW e PHOW-RGB. Com isso, também é feita uma comparação entre os descritores locais. Os métodos BOVW e pirâmide espacial foram utilizados para formar o descritor final da imagem. Seguindo trabalhos anteriores (Zou et al., 2012; Chen et al., 2011), foram utilizadas 30 imagens de cada classe para treinamento e o restante para os testes. Na etapa de classificação, a Máquina de Vetores de Suporte (do inglês - *Support Vector Machine* – SVM) foi utilizada (Wang, 2005). O SVM foi aplicado 10 vezes com imagens escolhidas aleatoriamente e, então foi calculada a média e o desvio padrão. Além disso, foram levados em consideração a taxa de classificação correta média (TCCM - Equação 4.1) e a medida-f (MF - Equação 4.2) como forma de avaliação dos métodos.

$$TCCM = \frac{VP+VN}{VP+VN+FP+FN} \tag{4.1}$$

$$PR = \frac{VP}{VP+FP}, \quad RE = \frac{VP}{VP+FN}, \quad MF = 2 * \frac{PR*RE}{PR+RE} \tag{4.2}$$

onde PR representa a precisão, RE a revocação, VP os verdadeiros positivos, FP os falsos positivos, FN os falsos negativos e VN os verdadeiros negativos. Os valores VP, VN, FP e FN são obtidos por meio da matriz de confusão. A matriz de confusão proposta por (Powers, 2011) é um tipo de tabela que permite a visualização do desempenho de um método de aprendizado, normalmente em um método supervisionado. A matriz de confusão é uma tabela com duas linhas e duas colunas que relata o número de verdadeiros positivos, falsos negativos, falsos positivos e verdadeiros negativos. Isso permite uma análise mais detalhada do que a mera proporção de suposições corretas (exatidão). Por exemplo, se houvesse 95 imagens de doenças e apenas 5 de não doenças no conjunto de dados, o classificador poderia facilmente ser influenciado na classificação de todas as amostras como doenças. A precisão geral seria de 95%, mas na prática o classificador teria uma taxa de reconhecimento de 100% para a classe doença, mas uma taxa de reconhecimento de 0% para a classe de não doença.

A Figura 4.2, apresenta um exemplo de matriz de confusão, onde cada linha da matriz representa os casos de uma classe real, enquanto as colunas representam as instâncias das classes preditas. No exemplo, foi computado uma matriz de confusão para um método de classificação que foi treinado para reconhecer uma determinada doença através do diagnóstico. Nesse exemplo, foi utilizado a amostra de 20 imagens onde há a presença da doença e 10 onde não há. Na matriz apresentada, de 20 imagens da doença, o sistema previu que 5 eram de não doença e das 10 de não doença, o sistema previu que 7 eram de não doença e 3 eram de doença. Desse modo, o classificador atingiu uma taxa de classificação correta média de 73.33%, uma precisão de 83.33%, revocação de 75% e a medida-f de 78.94%.

		Classe predita	
		Doença	Não Doença
Classe real	Doença	15	5
	Não Doença	3	7

VP 15 verdadeiros positivos (Doenças reais que foram corretamente classificadas)	**FN** 5 falsos negativos (Doenças que foram incorretamente classificadas como não doença)
FP 3 falsos positivos (Não doenças que foram incorretamente classificadas)	**VN** 7 verdadeiros negativos (Não doenças que foram corretamente classificadas)

Figura 4.2: Na imagem a esquerda é apresentado um exemplo de uma matriz de confusão, onde as linhas representam os casos de classe uma classe real e as colunas representam as classes preditas. Na imagem a direita são mostrados os valores de VP, FN, FP e VN, que são utilizados nas métricas da taxa de classificação correta média e a medida-f.

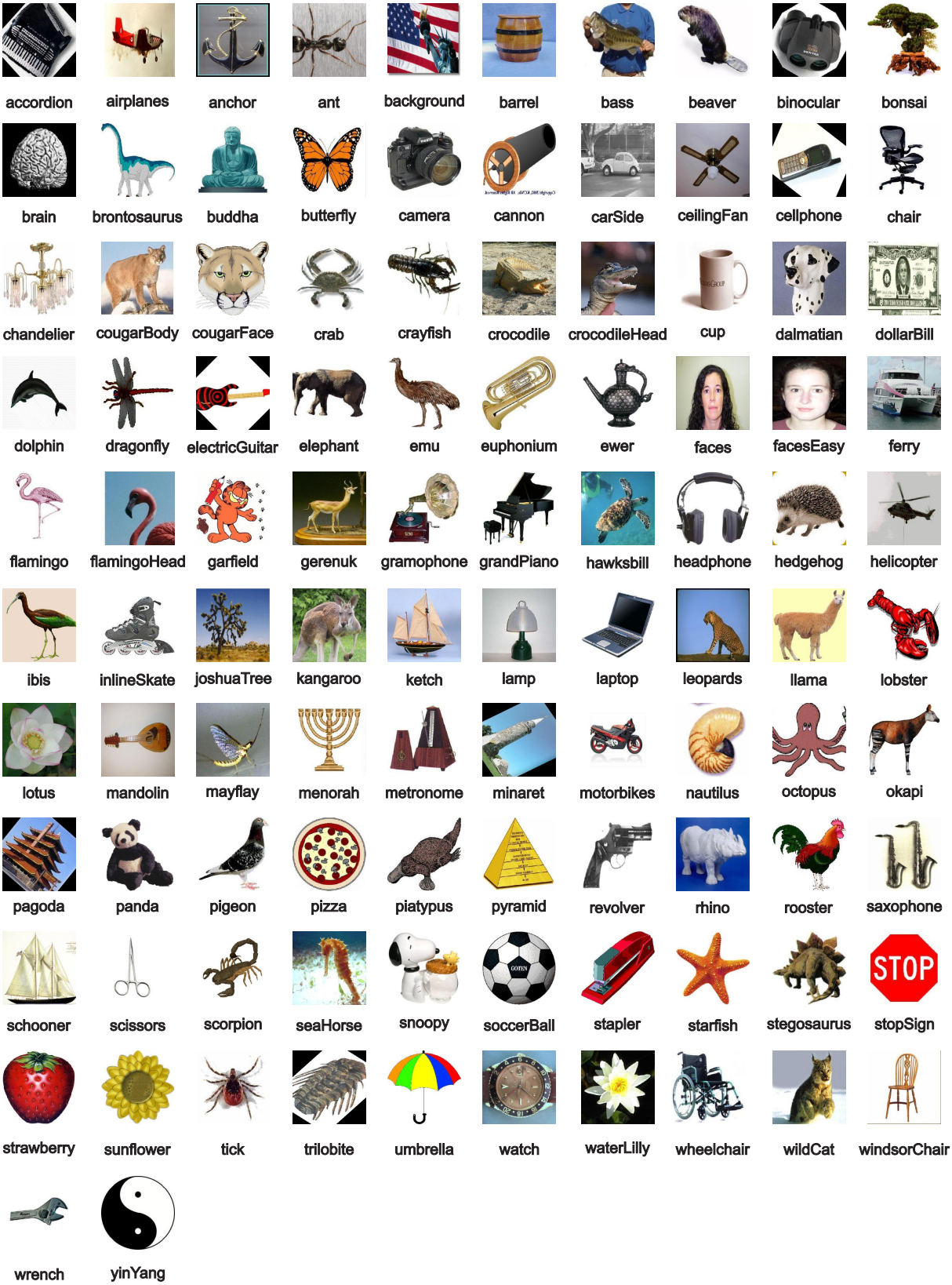


Figura 4.3: Exemplo das classes de imagens contidas na base Caltech-101. Na imagem estão presentes um exemplo de cada classe da base.

4.2 Resultados e Discussões

Nesta seção, são apresentados os resultados obtidos com os experimentos. A Seção 4.2.1 apresenta a análise de parâmetros do método proposto, enquanto que a Seção 4.2.2 mostra uma comparação entre o método proposto e os métodos BOVW e pirâmide espacial.

4.2.1 Análise de Parâmetros

Esta seção consiste em analisar o uso dos 2 parâmetros do método proposto: o tamanho do vocabulário de palavras visuais k, variando de 100 a 400 com incremento de 100 e o raio r de 0.1 a 1 com incremento de 0.1. Utilizamos uma maior variação para o raio, pelo fato da modelagem do grafo ser a principal contribuição do nosso método. O raio r é o parâmetro responsável pela modelagem do grafo, onde dois vértices são conectados caso a distância entre eles for menor que um determinado raio. O nível da pirâmide l é responsável por atribuir informações espaciais ao método proposto juntamente com a análise do grafo. Segundo Lazebnik et al., os melhores resultados foram obtidos com nível da pirâmide $l = 2$, por explorar mais as informações espaciais contidas nas imagens. Além disso, foi mostrado que é inviável utilizar o método com o nível da pirâmide l superior a 2, devido ao custo computacional exigido para executar o método. Desse modo, o método proposto foi utilizado com o nível da pirâmide ($l = 2$).

A Figura 4.4 apresenta 4 gráficos mostrando os resultados obtidos com o método proposto. O primeiro gráfico apresenta os resultados obtidos com o vocabulário de palavras visuais $k = 100$, o gráfico a direita na região superior utiliza $k = 200$. Os gráficos da região inferior utilizaram o vocabulário de palavras visuais $k = 300$ e $k = 400$, respectivamente. O eixo y dos gráficos representa a taxa de classificação correta média, enquanto que o eixo x representa o raio r. Em cada gráfico, são apresentados os resultados obtidos pelo método utilizando os descritores locais SIFT Denso, PHOW e PHOW-RGB, onde é possível fazer uma comparação de resultados entre os mesmos. Os gráficos apresentados na Figura 4.4 utilizam diferentes eixos para y, pelo fato das variações entre as taxas de classificação correta média de um mesmo k serem pequenas, e a diferença entre as TCCM, para diferentes k ser grande. Desse modo, se os gráficos fossem plotados com um mesmo eixo y para os diferentes valores de k, a análise para identificar qual é o melhor valor do raio r seria dificultada.

Em relação ao vocabulário de palavras visuais, utilizando o PHOW-RGB, o melhor resultado foi alcançado com $k = 400$, atingindo uma taxa de classificação correta média de 80.95% contra 80.55%, 79.66% e 77.92 % utilizando $k = 300$, 200 e 100, respectivamente. Desse modo, incrementando o tamanho do vocabulário de palavras visuais k, aumenta-se a taxa de classificação correta média. Da mesma forma, com o descritor local PHOW, o melhor resultado foi alcançado com $k = 400$, atingindo uma taxa de classificação correta média de 80.89%, enquanto $k = 300$, 200 e 100, atingiram 80.37%, 79.81% e 77.83%, respectivamente. Com o descritor SIFT Denso, a melhor taxa de classificação correta média também foi atingida com $k = 400$, onde foi alcançado 76.77%, seguido por $k = 300$, 200 e 100, com os resultados 76.64%, 76.52% e 74.84%, respectivamente. Com isso, podemos perceber que o descritor PHOW-RGB se mostrou superior aos descritores PHOW e SIFT Denso, independente do vocabulário de palavras visuais k, alcançando uma taxa de classificação correta média de 80.95%, enquanto o PHOW e SIFT Denso, alcançaram 80.89% e 76.77%, respectivamente. Considerando o intervalo [100, 400], a medida que k é incrementado podemos perceber que a taxa de classificação correta média aumenta, pois ao utilizar um maior número de palavras visuais, é feita uma melhor modelagem das características dos objetos. A comparação do tamanho do vocabulário de palavras visuais k para diferentes descritores é resumida na Tabela 4.1.

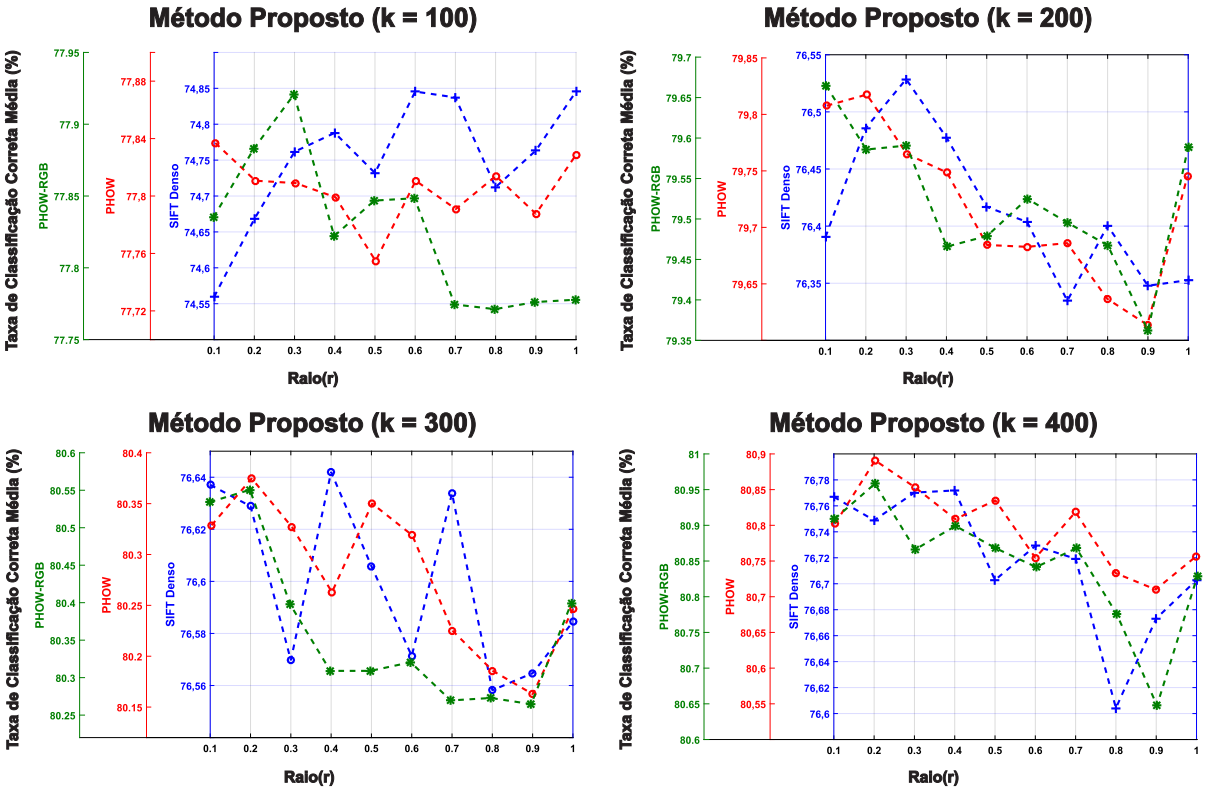


Figura 4.4: O gráfico apresenta no eixo y a taxa de classificação correta média do método proposto com diferentes raios e nível de pirâmide ($l = 2$). Os gráficos mostram resultados obtidos com o vocabulário de palavras visuais $k = 100$, 200, 300 e 400. O raio r, representado no eixo x foi variado de 0.1 a 1. Além disso, é apresentado os resultados alcançados pelos descritores locais SIFT Denso, PHOW e PHOW-RGB. Nos gráficos, os símbolos $*, +$, e $\circ$, são utilizados para representar os resultados obtidos com os descritores PHOW-RGB, PHOW e SIFT Denso, respectivamente.

		TCCM % (Desvio Padrão)				
k	SIFT Denso	Raio	PHOW	Raio	PHOW-RGB	Raio
100	74.84 (±0.43)	0.6	77.83 (±0.51)	0.1	77.92 (±0.39)	0.3
200	76.52 (±0.62)	0.3	79.81 (±0.38)	0.2	79.59 (±0.49)	0.3
300	76.64 (±0.46)	0.4	80.37 (±0.43)	0.2	80.55 (±0.42)	0.2
400	76.77 (±0.53)	0.4	80.89 (±0.46)	0.2	80.95 (±0.38)	0.2

Tabela 4.1: Resultados obtidos com o método proposto utilizando os descritores locais SIFT Denso, PHOW e PHOW-RGB. Na tabela é mostrado o tamanho do vocabulário de palavras k, a taxa de classificação correta média, o desvio padrão e o raio. Os resultados foram obtidos utilizando a base de imagens *Caltech-101*.

Também podemos notar nos gráficos da Figura 4.4, que a variação do raio r gera diferença nos resultados. Observando os resultados obtidos pelo método proposto, utilizando o descritor local PHOW-RGB, os melhores resultados foram obtidos com raio $r = 0.2$ com o vocabulário de palavras visuais $k = 400$. Com o descritor local PHOW, a maior taxa de classificação correta média foi atingida também com raio $r = 0.2$, com o vocabulário de palavras visuais $k = 400$. Por outro lado, com o descritor SIFT Denso os melhores resultados foram alcançados com o raio $r = 0.4$ com o vocabulário de palavras visuais $k = 400$. A comparação do tamanho do vocabulário de palavras visuais k e o raio r para diferentes descritores locais é resumida na Tabela 4.2.

Com isso, os melhores resultados foram obtidos com raios r entre 0.2 e 0.4. Os grafos modelados conectando apenas os pontos mais próximos ($r < 0.2$) acabam não incluindo informações locais suficientes para descrever o objeto, enquanto que os grafos que conectam

SIFT Denso										
k	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
100	74.55	75.66	74.76	74.78	74.73	**74.84**	74.83	74.71	74.76	**74.84**
200	76.39	76.48	**76.52**	76.47	76.41	76.40	76.33	76.40	76.34	76.35
300	76.63	76.62	76.56	**76.64**	76.60	76.57	76.63	76.55	76.56	76.58
400	76.76	76.74	**76.77**	**76.77**	76.70	76.72	76.71	76.60	76.67	76.70
PHOW										
100	**77.83**	77.81	77.80	77.79	77.75	77.81	77.79	77.81	77.78	77.82
200	79.80	**79.81**	79.76	79.74	79.68	79.68	79.68	79.63	79.61	79.74
300	80.32	**80.37**	80.32	80.26	80.35	80.31	80.22	80.18	80.16	80.24
400	80.80	**80.89**	80.85	80.80	80.83	80.75	80.81	80.73	80.71	80.75
PHOW-RGB										
k	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
100	77.83	77.88	**77.92**	77.82	77.84	77.84	77.77	77.77	77.77	77.77
200	79.66	79.58	**79.59**	79.46	79.47	79.52	79.49	79.46	79.36	79.58
300	80.53	**80.55**	80.39	80.30	80.30	80.32	80.26	80.27	80.26	80.39
400	80.90	**80.95**	80.86	80.89	80.86	80.84	80.86	80.77	80.64	80.82

Tabela 4.2: A Tabela apresenta a taxa de classificação correta média (TCCM) obtida pelos descritores SIFT Denso, PHOW e PHOW-RGB em relação ao tamanho do vocabulário de palavras k e o raio r.

pontos com distâncias grandes $(r > 0.5)$ não apresentam bons resultados, pois a topologia do grafo modelado inclui informações globais que não são relevantes no reconhecimento de objetos. Com os resultados obtidos, podemos definir como melhores parâmetros para o método proposto $l = 2, r = 0.2$ e $k = 400$, ao utilizar o PHOW e PHOW-RGB e $l = 2, r = 0.4$ e $k = 400$, ao utilizar o SIFT Denso.

4.2.2 Comparação com métodos da literatura

Primeiramente, foi avaliada a taxa de classificação correta média do método BOVW, que não inclui informações espacias em seu descritor final. Para realizar os experimentos, o tamanho do vocabulário de palavras visuais k foi variado de 100 a 1000, com incremento de 100, conforme apresentado no gráfico da Figura 4.5. Nesse gráfico, apresentamos a taxa de classificação correta média do método BOVW em relação ao tamanho do vocabulário de palavras visuais k, utilizando os descritores locais SIFT Denso, PHOW e PHOW-RGB.

Os resultados mostraram que quanto maior for o k, maior é a taxa de classificação correta média do método. Entretanto, a partir de $k = 600$ a taxa de classificação correta média é estabilizada, tendo pouco acréscimo na TCCM em cada incremento de k. Com $k = 100$, o método BOVW atingiu a taxa de classificação correta média de 52.74% com o PHOW-RGB, enquanto o PHOW e SIFT Denso alcançaram uma TCCM de 51.83% e 45.58%, respectivamente. O melhor resultado desta técnica, independente do descritor local, foi utilizando $k = 1000$, com essa configuração o método obteve uma taxa de classificação correta média de 70.66% com o descritor PHOW-RGB, 69.91% com o PHOW e 59.17% com o SIFT Denso.

Em seguida, avaliamos a taxa de classificação correta média do método pirâmide espacial. Nesse caso, utilizamos o mesmo nível da pirâmide indicado pelo autor $l = 2$ (Lazebnik et al., 2006). Porém, variamos o tamanho do vocabulário de palavras k de 100 a 400, pois segundo o autor é inviável utilizar $k > 400$ devido ao número grande de características que o método irá produzir e pouco acréscimo nos resultados de classificação (Lazebnik et al., 2006). A Tabela 4.3 apresenta os resultados alcançados pelo método pirâmide espacial, utilizando os descritores locais SIFT Denso, PHOW e PHOW-RGB. Além da taxa de classificação correta média, a tabela apresenta o tamanho do vocabulário de palavras k, o desvio padrão e o número de características (NC).

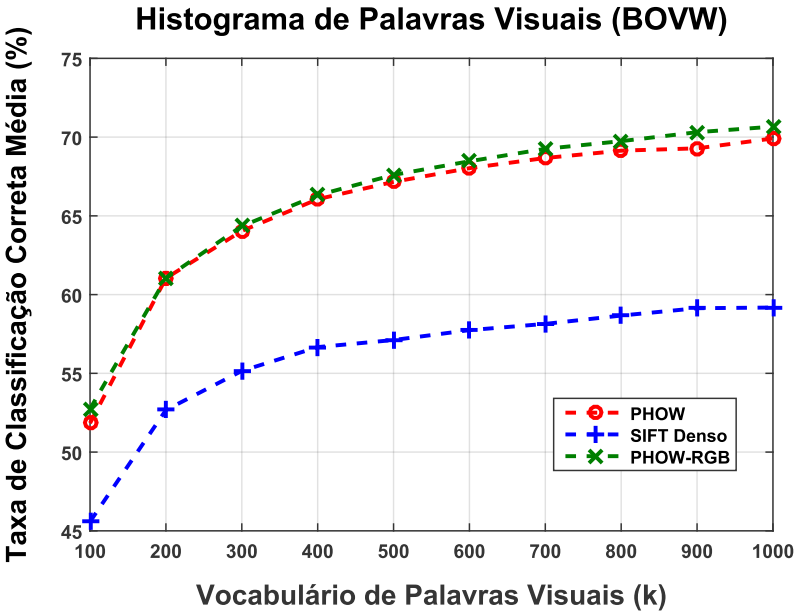


Figura 4.5: Taxa de classificação correta média utilizando o método BOVW com o tamanho do vocabulário de palavras visuais k variando de 100 a 1000. O eixo x é o tamanho do vocabulário de palavras visuais k e o eixo y representa a taxa de classificação correta média do BOVW utilizando os descritores locais SIFT Denso, PHOW e PHOW-RGB.

		TCCM % (Desvio Padrão)		
k	(NC)	SIFT Denso	PHOW	PHOW-RGB
100	2100	74.93 (±0.37)	77.60 (±0.53)	77.89 (±0.56)
200	4200	76.31 (±0.55)	79.55 (±0.53)	79.41 (±0.49)
300	6300	76.42 (±0.39)	80.10 (±0.44)	80.22 (±0.45)
400	8400	76.32 (±0.52)	80.56 (±0.35)	80.71 (±0.48)

Tabela 4.3: Resultados obtidos com o método pirâmide espacial utilizando os descritores locais SIFT Denso, PHOW e PHOW-RGB. Na tabela é mostrado o tamanho do vocabulário de palavras k, a taxa de classificação correta média, o desvio padrão e o número de características (NC) extraídas pelo método. Os resultados foram obtidos utilizando a base de imagens *Caltech-101*.

Com isso, percebemos que o melhores resultados foram alcançados utilizando o descritor local PHOW-RGB, que alcançou uma taxa de classificação correta média de 80.71%, seguido pelos descritores PHOW e SIFT Denso, que atingiram 80.56% e 76.42%, respectivamente. Também é possível perceber que incrementando o tamanho do vocabulário de palavras visuais k, a taxa de classificação correta média aumenta, exceto para o descritor SIFT Denso, que com $k = 400$ obteve uma taxa de classificação correta média inferior a $k = 300$. Contudo, percebe-se que a inclusão de informações espaciais com o método pirâmide espacial resultou em um ganho de 10.05% na taxa de classificação correta média em relação ao método BOVW.

A Tabela 4.4 apresenta um resultado comparativo entre os métodos BOVW, pirâmide espacial e o método proposto, utilizando a melhor taxa de classificação correta média obtida por cada método, com os descritores locais SIFT Denso, PHOW e PHOW-RGB. Além da taxa de classificação correta média e a medida-f, a tabela também apresenta o tamanho do vocabulário de palavras k, o raio r, o desvio padrão σ e o número de características (NC) que são importantes na comparação. Tanto o método proposto quanto o método pirâmide espacial utilizaram o nível da pirâmide $l = 2$ nos resultados apresentados.

SIFT Denso					
Método	k	NC	r	TCCM % (σ)	Medida-f (σ)
BOVW	1000	**1000**	-	59.17 (±0.60)	37.69 (±0.257)
Pirâmide Espacial	**300**	6300	-	76.42 (±0.39)	57.51 (±0.236)
Método Proposto	400	8400	0.2	**76.77 (**±0.53**)**	**57.94 (**±0.231**)**
PHOW					
Método	k	NC	r	TCCM % (σ)	Medida-f (σ)
BOVW	1000	**1000**	-	69.91 (±0.46)	50.28 (±0.247)
Pirâmide Espacial	**400**	8400	-	80.56 (±0.35)	64.64 (±0.218)
Método Proposto	**400**	8400	0.4	**80.89 (**±0.46**)**	**64.82 (**±0.219**)**
PHOW-RGB					
Método	k	NC	r	TCCM % (σ)	Medida-f (σ)
BOVW	1000	**1000**	-	70.66 (±0.52)	51.56 (±0.243)
Pirâmide Espacial	**400**	8400	-	80.71 (±0.48)	64.82 (±0.218)
Método Proposto	**400**	8400	0.4	**80.95 (**±0.38**)**	**64.96 (**±0.218**)**

Tabela 4.4: Comparação entre o método proposto, pirâmide espacial e BOVW, utilizando a base *Caltech-101*. A Tabela apresenta o tamanho do vocabulário de palavras k, o número de características (NC), o raio r, a taxa de classificação correta média (TCCM), o desvio padrão (σ) e a medida-f que são interessantes para a comparação.

Com os resultados, podemos perceber que a inclusão de informação espacial de fato apresenta uma melhoria significativa na taxa de classificação correta média. Porém, o número de características do método proposto e pirâmide espacial são superiores ao do BOVW fazendo com que eles sejam mais custosos na etapa de classificação. Além disso, podemos apontar um ganho do método proposto em relação ao pirâmide espacial, pelo fato de explorar informações espaciais em cada sub-região da imagem. Também, podemos concluir que a utilização do descritor local influencia no resultado final, pois utilizando o descritor local PHOW-RGB, o método proposto e pirâmide espacial tiveram um aumento significativo na taxa de classificação correta média em relação aos demais descritores. Em relação ao parâmetro raio r, o método proposto apresentou os melhores resultados com $r = 0.2$ e 0.4. Com isso, podemos concluir que a topologia dos grafos modelados com apenas as palavras mais próximas espacialmente, resultam em características melhores para o descritor final da imagem e, consequentemente, em uma taxa de classificação correta média melhor para o método proposto.

A Tabela 4.5 apresenta as sete classes que obtiveram as melhores taxas de classificação correta média e as sete piores em relação ao método proposto, utilizando o descritor PHOW-RGB. Além disso, é apresentado o número de imagens (NI) de cada classe. Com os resultados, podemos perceber que o número de imagens não influência na TCCM, mas sim a variação intraclasse das imagens. O método proposto obteve os melhores resultados com as classes "Leopards", "Accordion", "Minaret", "CarSide", "Trilobite", "DollarBill" e "Facescom", com uma taxa de classificação correta média de 100%, 100%, 100%, 99.89%, 99.64%, 99.55% e 99.11%, respectivamente. O método proposto foi superior ao pirâmide espacial em 3 classes, empatou em 3 e perdeu em apenas 1. O método BOVW, foi inferior aos métodos proposto e pirâmide espacial em todas as 7 classes apresentadas na Tabela 4.5.

Apesar disso, o método proposto alcançou os piores resultados com as classes "BGGoogle", "Crocodile", "CougarBody", "Ant", "Beaver", "Octopus" e "Cannon" com TCCM de 22.15%, 28.50%, 30.59%, 33.33%, 36.88%, 38% e 51.54%, nessa ordem. Essas classes possuem uma grande variação intraclasse, o que as tornam mais complicadas de serem reconhecidas pelos métodos, resultando em uma TCCM baixa. Através da Tabela 4.5, podemos perceber que para as classes mais complicadas de serem reconhecidas o método proposto se mostrou superior ao pirâmide espacial e BOVW. O método proposto alcançou resultados melhores em 4 classes, empatou em 1 e perdeu em 2. Dessa forma, o método proposto se mostrou superior aos

demais métodos comparados, em relação as classes com os melhores e piores resultados. A Tabela com os resultados obtidos pelo método proposto, pirâmide espacial e BOVW, utilizando os descritores locais SIFT Denso e PHOW, são apresentadas no Apêndice A.

Melhores resultados				
		PHOW-RGB		
Classe	NI	BOVW	Pirâmide Espacial	Método Proposto
Leopards	200	99.88	**100**	**100**
Accordion	55	96.40	**100**	**100**
Minaret	76	98.91	**100**	**100**
CarSide	123	99.57	**100**	99.89
Trilobite	86	98.04	99.46	**99.64**
DollarBill	52	93.64	99.09	**99.55**
Faces	435	85.56	98.49	**99.11**
Piores resultados				
		PHOW-RGB		
Classe	NI	BOVW	Pirâmide Espacial	Método Proposto
Cannon	43	40.77	50	**51.54**
Octopus	35	30	**42**	38
Beaver	46	24.37	32.50	**36.88**
Ant	42	22.50	**33.33**	**33.33**
CougarBody	47	19.41	**35.29**	30.59
Crocodile	50	28	27	**28.50**
BGGoogle	467	18.44	20.55	**22.15**

Tabela 4.5: Resultados obtidos pelos métodos BOVW, pirâmide espacial e método proposto. Na tabela é mostrada a taxa de classificação correta média obtida pelos três métodos levando em consideração as classes com os melhores e piores resultados. Além disso, também é mostrado o número de imagens (NI) das classes apresentadas.

As Figura 4.6 apresenta os resultados obtidos pelo método proposto, pirâmide espacial e BOVW em relação a cada classe da base Caltech-101. O gráfico mostra que o método proposto foi superior em 48% das classes, empatou em 16% e perdeu em 36%, comparado com o pirâmide espacial. Em comparação com o BOVW, o método proposto obteve uma taxa de classificação correta média superior em 97% das classes, empatou em 1% e perdeu em 2% das classes. Com os resultados, é possível perceber que em classes onde há distribuição uniforme dos pontos de interesse, a informação espacial extraída pelo método é desnecessária e, com isso, o método proposto se mostrou inferior aos demais métodos comparados. No Apêndice A são apresentados os resultados obtidos pelo método proposto, pirâmide espacial e BOVW, utilizando os descritores locais SIFT Denso e PHOW, em relação a cada classe da base Caltech-101.

Também podemos perceber que, utilizando o descritor local PHOW-RGB, o método proposto alcançou uma TCCM superior a 90% em 23 classes, enquanto o método pirâmide espacial obteve TCCM superior a 90% em 21 classes e o BOVW em 11 classes. Os resultados ressaltam a superioridade do método proposto em relação aos demais métodos comparados, onde foi mostrado que o método proposto obteve um resultado superior em mais classes que os métodos pirâmide espacial e BOVW. Além disso, foi mostrado que o método proposto obteve mais classes com uma taxa de classificação correta média superior a 90% que os métodos comparados.

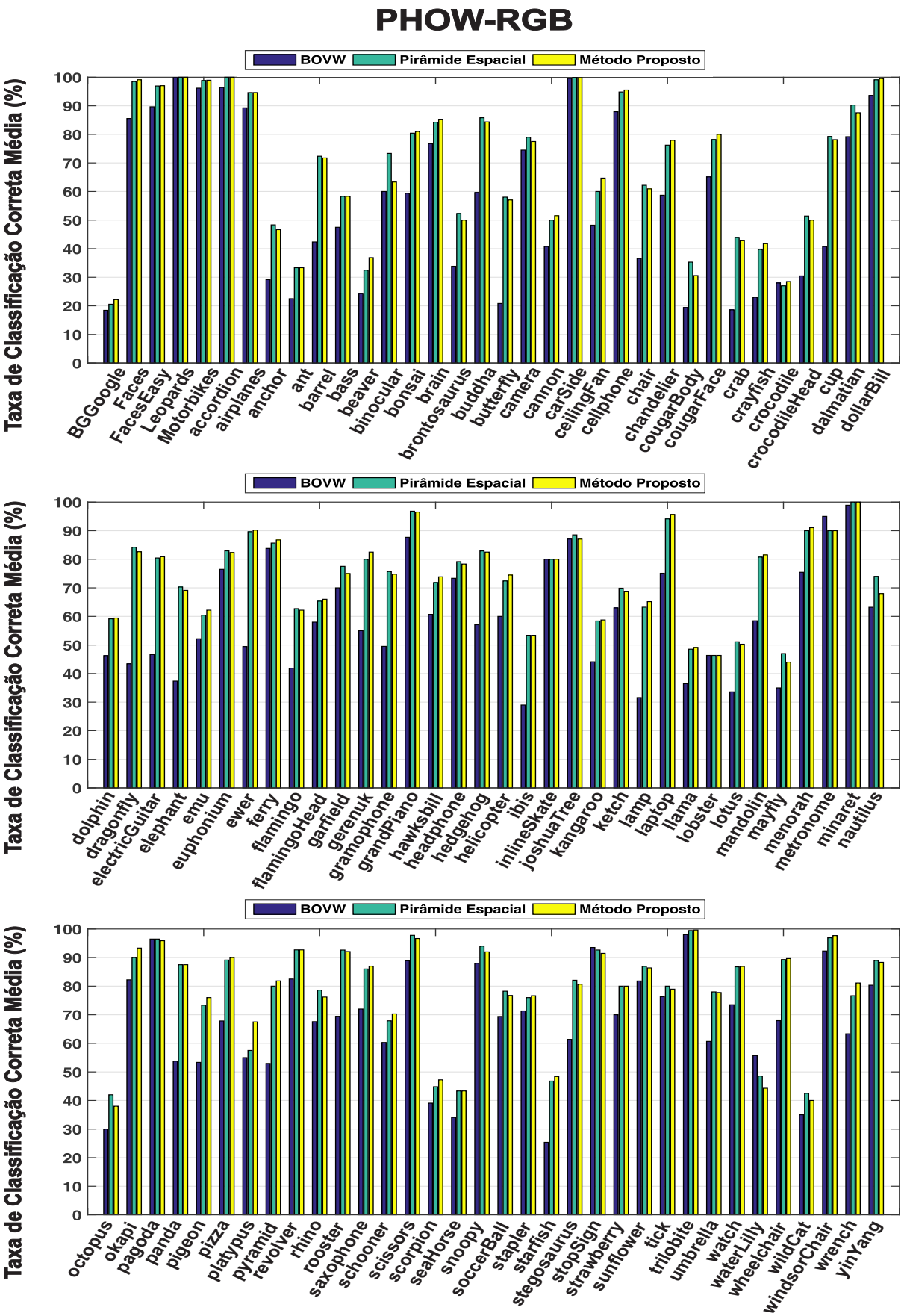


Figura 4.6: Comparação de resultados entre os métodos BOVW, Pirâmide Espacial e método proposto, em relação a cada classe da base Caltech-101. A imagem mostra o melhor resultado de cada método como mostrado na Tabela 4.4, utilizando o PHOW-RGB como descritor local.

CAPÍTULO

5

Aplicação: Reconhecimento de Vagas de Estacionamento

Devido as significativas mudanças sociais e ao grande aumento da frota de automóveis nas cidades, a necessidade de se encontrar vagas para estacionamento se tornou um problema em potencial (Falcão et al., 2013). Com os avanços da visão computacional, é possível realizar o reconhecimento de vagas de estacionamento de modo automatizado, auxiliando os motoristas a encontrarem estacionamentos disponíveis de maneira mais rápida (Amato et al., 2016, 2017). Neste capítulo, são apresentados resultados da utilização do método proposto em uma aplicação de reconhecimento de vagas de estacionamento, que é uma boa forma de avaliar o método proposto, por se tratar de uma aplicação real. Na Seção 5.1 é mostrado o protocolo experimental, onde são descritos os parâmetros escolhidos para o método proposto, bem como, a base de imagens utilizada. A Seção 5.2, apresenta os resultados obtidos e uma discussão sobre eles.

5.1 Protocolo Experimental

Para realizar os experimentos, foi utilizada a base de imagens *PKLot* proposta por De Almeida et al. (2015). A *PKLot* contém 12.417 imagens de estacionamentos e 695.899 imagens de vagas de estacionamento segmentadas a partir delas, que foram manualmente verificadas e rotuladas de acordo com a situação (disponível ou ocupada), como ilustrado na Figura 5.1. As vagas de estacionamento foram capturadas de diferentes estacionamentos sob condições climáticas variadas (ensolarado, nublado ou chuvoso) e sem controle de iluminação, como mostrado na Figura 5.2. As imagens foram capturadas por uma câmera full HD de baixo custo (Microsoft LifeCam, HD-5000) posicionada no topo de um prédio para minimizar possíveis oclusões entre veículos vizinhos. Todas as imagens foram adquiridas nos parques de estacionamento da Universidade Federal do Paraná (UFPR) e da Pontifícia Universidade Católica do Paraná (PUCPR), ambas localizadas em Curitiba, Brasil.

Nos experimentos, foram utilizadas 200 mil imagens escolhidas de forma aleatória, onde 2 mil foram utilizadas para treinamento e o restante para os testes (99 mil imagens de va-

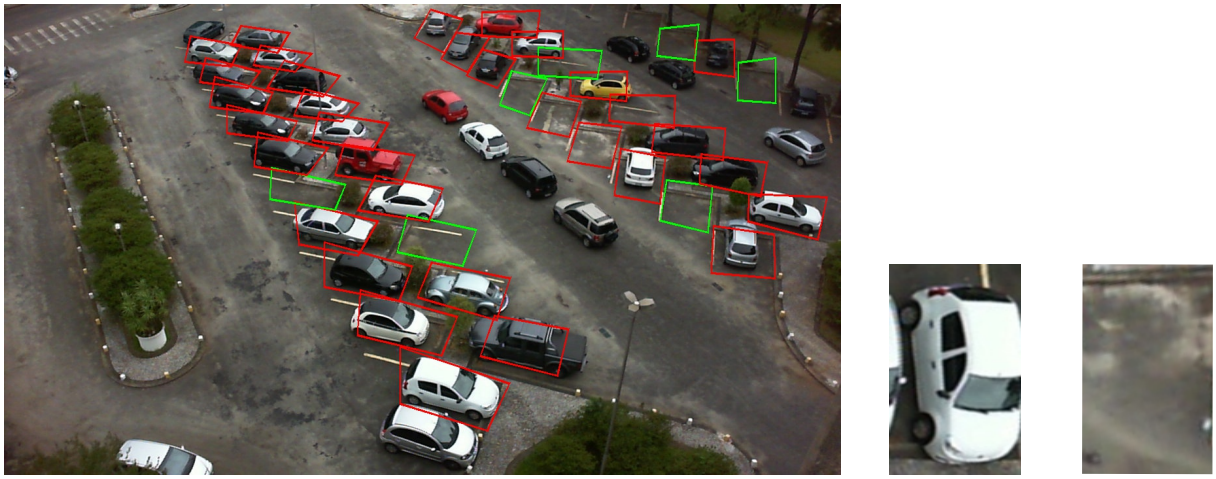


Figura 5.1: Exemplo da segmentação feita pelos autores nas imagens de estacionamento. Na primeira imagem, temos 40 espaços delimitados, na segunda o recorte de uma vaga ocupada e por último o recorte de uma vaga disponível.

Figura 5.2: Exemplos de imagens da base *PKLot* capturadas sob condições climáticas diferentes: na primeira coluna nublado, na segunda chuvoso e na última ensolarado.

gas de estacionamento disponíveis e 99 mil imagens de vagas de estacionamento ocupadas). Desse modo, podemos perceber que o conjunto de treinamento é muito menor que o conjunto de testes, tornando o reconhecimento um desafio. O método proposto foi utilizado com os parâmetros obtidos na análise realizada da Seção 4.2.1. Desse modo, foram utilizados os descritores locais SIFT Denso com raio $r = 0.4$, PHOW e PHOW-RGB com raio $r = 0.2$. Além disso, foi utilizado o vocabulário de palavras visuais $k = 400$ em todos os testes. O SVM foi utilizado na etapa de classificação, onde foi aplicado 10 vezes, e então foi calculada a média e o desvio padrão. Desse modo, foram utilizadas como métricas a taxa de classificação correta média (*TCCM* - Equação 4.1) e a medida-f (*MF* - Equação 4.2) que são importantes para a avaliação do método proposto.

5.2 Resultados e Discussões

A Tabela 5.1 apresenta os resultados obtidos, em que é possível perceber que o descritor local PHOW se mostrou superior aos demais descritores, alcançando uma taxa de classificação correta média de 99.81%, contra 99.78% e 99.26% dos descritores locais PHOW-RGB e SIFT Denso, respectivamente. A vantagem de se utilizar o PHOW, pode ser explicada pelo fato da informação de cor inclusa pelo PHOW-RGB não ser relevante nesse tipo aplicação, pois o objetivo da aplicação é reconhecer se a vaga de estacionamento está disponível ou ocupada, independente da cor do automóvel. Além disso, é possível notar que independente do descritor local utilizado, os resultados foram ótimos por alcançarem uma *TCCM* maior que 99%.

Descritor	k	raio	TCCM % (Desvio Padrão)	Medida-f
SIFT Denso	400	0.2	99.26 ($\pm$0.11)	99.26 ($\pm$0.00)
PHOW	400	0.4	99.81 ($\pm$0.07)	99.82 ($\pm$0.00)
PHOW-RGB	400	0.4	99.78 ($\pm$0.06)	99.78 ($\pm$0.00)

Tabela 5.1: Resultados obtidos com o método proposto no reconhecimento de vagas de estacionamento. Nos experimentos, foram utilizados os descritores locais SIFT Denso, PHOW e PHOW-RGB. A tabela mostra o tamanho do vocabulário de palavras visuais k, o raio r, a taxa de classificação correta média *TCCM*, o desvio padrão e a medida-f. Os resultados foram obtidos utilizando a base de imagens *PKLot*.

A Figura 5.3 apresenta as matrizes de confusão obtidas pelo método proposto utilizando os descritores locais SIFT Denso, PHOW e PHOW-RGB. Com os resultados, podemos perceber que apesar do método PHOW ter alcançado uma maior taxa de classificação correta média, o descritor PHOW-RGB confundiu menos as vagas de estacionamento ocupadas, ou seja, em apenas $0,03\%$ dos testes o sistema confundiu as vagas disponíveis com vagas ocupadas, contra $0,04\%$ de confusão entre as imagens ocupadas do descritor local PHOW. Além disso, é notável que para todos os descritores locais analisados, houve uma maior confusão entre as vagas disponíveis, que foram consideradas de forma incorreta como ocupadas, com $1,21\%$, $0,41\%$ e $0,33\%$ para os descritores SIFT Denso, PHOW-RGB e PHOW, respectivamente.

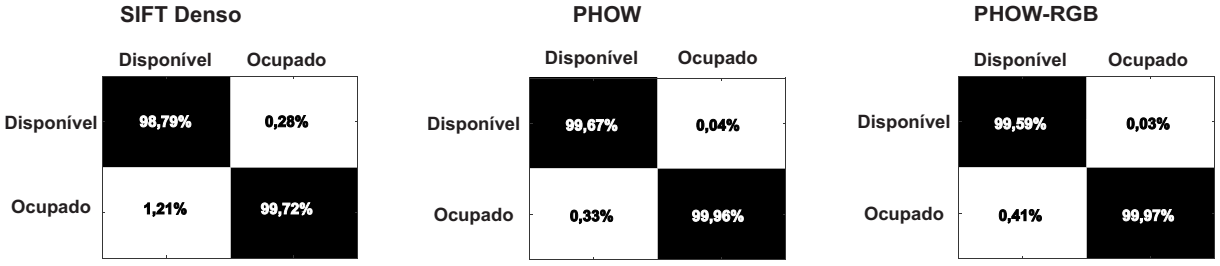


Figura 5.3: Matrizes de confusão da aplicação do método proposto com os descritores locais SIFT Denso, PHOW e PHOW-RGB. Nos testes foram utilizados 99 mil imagens de estacionamentos disponíveis e 99 mil imagens de estacionamentos ocupados.

A Figura 5.4 apresenta um exemplo da aplicação do método proposto no problema do reconhecimento de vagas de estacionamento. As imagens do exemplo são de diferentes estacionamentos que estão contidas na base *PKLot*, onde a localização das vagas foram determinadas manualmente pelo autor da base (De Almeida et al., 2015). Desse modo, para realizar o reconhecimento, as regiões da imagem onde existem estacionamento foram segmentadas, o classificador então reconhece como uma vaga de estacionamento disponível ou ocupada, e por fim, faz uma marcação colorida entorno da vaga, onde a cor verde representa que a vaga está disponível e a cor vermelha representa que a vaga está ocupada. Com isso, é possível utilizar a aplicação para auxiliar os motoristas a encontrarem vagas de estacionamento disponível de

modo mais rápido, evitando congestionamento dentro do estacionamento.

Figura 5.4: Exemplo da aplicação do método proposto no reconhecimento de vagas de estacionamento. No exemplo é mostrada imagens da base *PKLot*, onde a localização das vagas foram rotuladas manualmente pelo autor De Almeida et al. (2015). A partir da localização de cada vaga, a imagem da região é segmentada e então reconhecida como disponível ou ocupada.

CAPÍTULO

6

Conclusões

Neste capítulo, são apresentadas as conclusões obtidas com este trabalho. Na Seção 6.1 é realizado um paralelo entre os objetivos desta dissertação e os principais resultados alcançados. Na Seção 6.2 são apresentadas algumas direções e possibilidades para os trabalhos futuros.

6.1 Resumo dos Objetivos e Principais Resultados

Este trabalho apresentou uma abordagem para tornar o método histograma de palavras visuais mais preciso, onde adicionamos características locais por meio da combinação do método pirâmide espacial e a modelagem das palavras visuais em grafos. Por meio dos resultados é possível perceber um ganho na taxa de classificação correta média entre os métodos comparados que utilizam informações espaciais em seu descritor e os que não utilizam. Os resultados mostram que a utilização de grafos para extrair características é uma abordagem capaz de alcançar resultados mais vantajosos que a técnica de pirâmide espacial, pelo fato de explorar informações espaciais dentro de cada sub-região em que a imagem é dividida.

Os experimentos mostraram que para a base de imagens Caltech-101, uma base para reconhecimento de objetos, o melhor valor de r é 0.2 quando utilizado os descritores locais PHOW e PHOW-RGB, enquanto que com o descritor SIFT Denso, o melhor valor para r é 0.4. Desse modo, podemos perceber que os grafos modelados conectando apenas os pontos mais próximos $(r < 0.2)$ não incluem informações locais suficientes para descrever o objeto, enquanto que os grafos que conectam pontos com distâncias grandes $(r > 0.5)$ não apresentam bons resultados, pois a topologia do grafo modelado inclui informações globais que não são relevantes no reconhecimento de objetos.

Além disso, é possível notar que independente do descritor local, quanto maior o tamanho do vocabulário de palavras k, melhor é o resultado. Porém, a partir de $k > 400$ o custo computacional cresce de forma desproporcional à taxa de classificação correta média. Com isso, o melhor resultado foi obtido com o vocabulário de palavras visuais $k = 400$. Contudo, em classes onde há distribuição uniforme dos pontos de interesse, o método proposto apresentou resultados inferiores aos métodos comparados, pelo fato da informação espacial incluída pelo método ser desnecessária nesses casos. Por fim, foram realizados experimentos com o

método proposto em uma aplicação real, em que o mesmo foi aplicado no problema do reconhecimento de vagas de estacionamento. Os experimentos foram realizados utilizando a base de imagens *PKLot*, no qual o método proposto se mostrou eficiente para a solução desse problema em potencial, alcançando uma taxa de classificação correta média de 99.81%.

6.2 Trabalhos Futuros

Como continuação do trabalho, pretendemos otimizar o método para extração de características, melhorando o modo da modelagem dos grafos. Por exemplo, ao invés de modelar um grafo para cada palavra visual, construir um único grafo, extraindo características da topologia desse grafo, fazendo com que o método tenha um melhor desempenho. Em nosso método, foi utilizado somente a medida do grau médio para descrever os grafos, desse modo, pretendemos melhorar os resultados, extraindo outras medidas da topologia do grafo (e.g., energia, entropia, etc.). Além disso, pretendemos incluir uma comparação com outros métodos presentes na literatura, como por exemplo, métodos que incluem informação espacial através da correlação de palavras visuais, utilizar novas bases para validação do método proposto e os demais métodos, com a pretensão de explorar mais as invariâncias que o método proposto apresenta, como a variação à rotação e escala.

Atualmente, as redes neurais convolucionais (do inglês, *Convolutional Neural Networks*), também conhecida como aprendizado profundo (do inglês, *Deep learning*) tem apresentado resultados promissores em diversas áreas de visão computacional, tais como reconhecimento de faces (Liu et al., 2015), objetos (Eitel et al., 2015) e cenas (Wang et al., 2017). Desse modo, pretendemos incluir a técnica de modelagem de grafos e extração de características do método proposto nas camadas criadas pelo aprendizado profundo, com o objetivo de alcançar melhores resultados no problema do reconhecimento de objetos da base de imagens Caltech-101.

Referências Bibliográficas

Amato, G., Carrara, F., Falchi, F., Gennaro, C., Meghini, C., e Vairo, C. (2017). Deep learning for decentralized parking lot occupancy detection. *Expert Systems with Applications*, 72:327–334. Citado na página 40.

Amato, G., Carrara, F., Falchi, F., Gennaro, C., e Vairo, C. (2016). Car parking occupancy detection using smart camera networks and deep learning. In *Computers and Communication (ISCC), 2016 IEEE Symposium on*, páginas 1212–1217. IEEE. Citado na página 40.

Bay, H., Tuytelaars, T., e Van Gool, L. (2006). Surf: Speeded up robust features. In *Computer vision–ECCV 2006*, páginas 404–417. Springer. Citado nas páginas 2 e 14.

Bolovinou, A., Pratikakis, I., e Perantonis, S. (2013). Bag of spatio-visual words for context inference in scene classification. *Pattern Recognition*, 46(3):1039–1053. Citado nas páginas 5 e 14.

Bosch, A., Zisserman, A., e Muoz, X. (2007). Image classification using random forests and ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, páginas 1–8. IEEE. Citado nas páginas 2 e 13.

Brown, M. e Lowe, D. G. (2002). Invariant features from interest point groups. In *BMVC*, number s 1. Citado na página 9.

Brown, M. e Susstrunk, S. (2011). Multi-spectral sift for scene category recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, páginas 177–184. IEEE. Citado nas páginas 1 e 7.

Chen, B., Polatkan, G., Sapiro, G., Carin, L., e Dunson, D. B. (2011). The hierarchical beta process for convolutional factor analysis and deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, páginas 361–368. Citado na página 31.

Csurka, G., Dance, C., Fan, L., Willamowski, J., e Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, páginas 1–2. Prague. Citado nas páginas 1 e 14.

Dalal, N. e Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, páginas 886–893. IEEE. Citado na página 2.

De Almeida, P. R., Oliveira, L. S., Britto, A. S., Silva, E. J., e Koerich, A. L. (2015). Pklot–a robust dataset for parking lot classification. *Expert Systems with Applications*, 42(11):4937–4949. Citado nas páginas xvi, 4, 40, 42, e 43.

Dorkó, G. e Schmid, C. (2003). Selection of scale-invariant parts for object class recognition. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, páginas 634–639. IEEE. Citado na página 11.

Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., e Burgard, W. (2015). Multimodal deep learning for robust rgb-d object recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, páginas 681–687. IEEE. Citado na página 45.

Falcão, H. S., Lovato, A. V., dos Santos, A. F., Lucas Santos de Oliveira, R., Guimarães, M., e Santana, M. (2013). Classificação de vagas de estacionamento com utilização de rede perceptron multicamadas. *Revista de Sistemas de Informação da FSMA*, (12):41–48. Citado nas páginas 3 e 40.

Fei-Fei, L., Fergus, R., e Perona, P. (2007). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70. Citado na página 30.

Goldbarg, M. (2012). *Grafos: Conceitos, algoritmos e aplicações*. Elsevier Brasil. Citado na página 18.

Gonçalves, W. N. (2013). *Análise de texturas estáticas e dinâmicas e suas aplicações em biologia e nanotecnologia*. PhD thesis, Universidade de São Paulo. Citado na página 3.

Han, K., Rezende, R. S., Ham, B., Wong, K.-Y. K., Cho, M., Schmid, C., e Ponce, J. (2017). Scnet: Learning semantic correspondence. *arXiv preprint arXiv:1705.04043*. Citado na página 3.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features*. Springer. Citado na página 14.

Ke, Y. e Sukthankar, R. (2004). Pca-sift: A more distinctive representation for local image descriptors. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, páginas 506–513. IEEE Computer Society. Citado na página 11.

Khan, R., Barat, C., Muselet, D., e Ducottet, C. (2015). Spatial histograms of soft pairwise similar patches to improve the bag-of-visual-words model. *Computer Vision and Image Understanding*, 132:102–112. Citado na página 5.

Krapac, J., Verbeek, J., e Jurie, F. (2011). Modeling spatial layout with fisher vectors for image categorization. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, páginas 1487–1494. IEEE. Citado na página 5.

Lara, A. C. (2013). *Descritor de bordas e quantização espacial flexível aplicados a categorização de objetos*. PhD thesis, Universidade de São Paulo. Citado nas páginas 10 e 14.

Lazebnik, S., Schmid, C., e Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, páginas 2169–2178. IEEE. Citado nas páginas 2, 3, 4, 17, 33, e 35.

Liu, C., Yuen, J., e Torralba, A. (2016). Sift flow: Dense correspondence across scenes and its applications. In *Dense Image Correspondences for Computer Vision*, páginas 15–49. Springer. Citado nas páginas 1, 2, 11, e 24.

Liu, D., Hua, G., Viola, P., e Chen, T. (2008). Integrated feature selection and higher-order spatial feature extraction for object categorization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, páginas 1–8. IEEE. Citado na página 4.

Liu, Z., Luo, P., Wang, X., e Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, páginas 3730–3738. Citado na página 45.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, páginas 1150–1157. Ieee. Citado na página 8.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110. Citado nas páginas 1, 7, 8, 10, 11, 14, e 24.

Mikolajczyk, K. e Schmid, C. (2005). A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630. Citado na página 14.

Morioka, N. e Satoh, S. (2010). Building compact local pairwise codebook with joint feature space clustering. In *Computer Vision–ECCV 2010*, páginas 692–705. Springer. Citado na página 4.

Ojala, T., Pietikainen, M., e Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987. Citado na página 2.

Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. Citado na página 31.

Quack, T., Ferrari, V., Leibe, B., e Gool, L. V. (2007). Efficient mining of frequent and distinctive feature configurations. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, páginas 1–8. IEEE. Citado na página 4.

Rosa, G., Papa, J., Costa, K., Passos, L., Pereira, C., e Yang, X.-S. (2016). Learning parameters in deep belief networks through firefly algorithm. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, páginas 138–149. Springer. Citado na página 3.

Sánchez, J., Perronnin, F., e De Campos, T. (2012). Modeling the spatial layout of images beyond spatial pyramids. *Pattern Recognition Letters*, 33(16):2216–2223. Citado na página 5.

Savarese, S., Winn, J., e Criminisi, A. (2006). Discriminative object class models of appearance and shape by correlatons. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, páginas 2033–2040. IEEE. Citado na página 4.

Se, S., Lowe, D., e Little, J. (2002). Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *The international Journal of robotics Research*, 21(8):735–758. Citado na página 11.

Shah, S. A. A., Bennamoun, M., e Boussaid, F. (2016). Iterative deep learning for image set based face and object recognition. *Neurocomputing*, 174:866–874. Citado na página 1.

Vedaldi, A. e Fulkerson, B. (2010). Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the international conference on Multimedia*, páginas 1469–1472. ACM. Citado na página 8.

Wang, L. (2005). *Support vector machines: theory and applications*, volume 177. Springer Science & Business Media. Citado na página 31.

Wang, X., Lu, L., Shin, H.-c., Kim, L., Bagheri, M., Nogues, I., Yao, J., e Summers, R. M. (2017). Unsupervised joint mining of deep features and image labels for large-scale radiology image categorization and scene recognition. *arXiv preprint arXiv:1701.06599*. Citado na página 45.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37. Citado nas páginas 2 e 15.

Zagoris, K., Pratikakis, I., Antonacopoulos, A., Gatos, B., e Papamarkos, N. (2014). Distinction between handwritten and machine-printed text based on the bag of visual words model. *Pattern Recognition*, 47(3):1051–1062. Citado na página 14.

Zou, W., Zhu, S., Yu, K., e Ng, A. Y. (2012). Deep learning of invariant features via simulated fixations in video. In *Advances in neural information processing systems*, páginas 3212–3220. Citado na página 31.

APÊNDICE

A

Resultados Complementares

Conforme indicado no Capítulo 4, neste Apêndice são apresentados resultados complementares dos experimentos realizados. Primeiramente, são apresentados os resultados obtidos pelo método proposto (MP), pirâmide espacial e BOVW, utilizando os descritores locais SIFT Denso e PHOW, em relação as classes que obtiveram os melhores e piores resultados da base Caltech-101. Em seguida, são mostrados os resultados obtidos pelos métodos analisados, utilizando os descritores locais SIFT Denso e PHOW, em relação a todas as classes da base.

A Tabela 4.5 apresenta as sete classes que obtiveram as melhores taxas de classificação correta média e as sete piores em relação ao método proposto, utilizando os descritores SIFT Denso, PHOW e PHOW-RGB. Além disso, é apresentado o número de imagens (NI) de cada classe. Com os resultados, podemos perceber que o número de imagens não influência na TCCM, mas sim a variação intraclasse das imagens. O método proposto obteve os melhores resultados com as classes "Leopards", "Accordion", "Minaret", "CarSide", "Trilobite", "DollarBill" e "Facescom", com uma taxa de classificação correta média de 100%, 100%, 100%, 99.89%, 99.64%, 99.55% e 99.11%, respectivamente. Com o descritor local PHOW-RGB o método proposto foi superior em 3 classes, empatou em 3 e perdeu em apenas 1. Com o PHOW, o método proposto obteve TCCM superior ao pirâmide espacial em 2 classes, empatou em 4 e perdeu em 1. Por outro lado, com o SIFT Denso, o método proposto foi melhor em 2 classes, empatou em 3 e perdeu em 2. O método BOVW, foi inferior aos métodos proposto e pirâmide espacial em todas as 7 classes apresentadas na Tabela 4.5.

Apesar disso, o método proposto alcançou os piores resultados com as classes "BGGoogle", "Crocodile", "CougarBody", "Ant", "Beaver", "Octopus" e "Cannon" com TCCM de 22.15%, 28.50%, 30.59%, 33.33%, 36.88%, 38% e 51.54%, nessa ordem. Essas classes possuem uma grande variação intraclasse, o que as tornam mais complicadas de serem reconhecidas pelos métodos, resultando em uma TCCM baixa. Através da Tabela 4.5, podemos perceber que para as classes mais complicadas de serem reconhecidas o método proposto se mostrou superior ao pirâmide espacial. Com o descritor local PHOW-RGB, o método proposto alcançou resultados melhores em 4 classes, empatou em 1 e perdeu em 2. Com o SIFT Denso, o método proposto foi superior em 4 classes e inferior em 3. Já com o PHOW, a diferença foi ainda maior, em que o MP obteve uma TCCM maior em 5 classes, contra 2 classes do pirâmide espacial. Dessa

forma, o método proposto se mostrou superior aos demais métodos comparados, em relação as classes com os melhores e piores resultados.

Melhores resultados										
		SIFT Denso			**PHOW**			**PHOW-RGB**		
Classe	NI	BOVW	SP	MP	BOVW	SP	MP	BOVW	SP	MP
Leopards	200	95.71	**100**	**100**	99.82	**100**	**100**	99.88	**100**	**100**
Accordion	55	96.80	**99.60**	**99.60**	94.40	**100**	**100**	96.40	**100**	**100**
Minaret	76	98.91	**99.78**	99.57	99.57	**100**	**100**	98.91	**100**	**100**
CarSide	123	93.55	**99.68**	**99.68**	98.28	**100**	**100**	99.57	**100**	99.89
Trilobite	86	96.96	**99.11**	98.57	96.25	**99.82**	99.64	98.04	99.46	**99.64**
DollarBill	52	89.09	96.36	**97.73**	93.64	**98.64**	97.73	93.64	99.09	**99.55**
Faces	435	91.01	98.25	**99.04**	85.43	98.67	**99.23**	85.56	98.49	**99.11**
Piores resultados										
		SIFT Denso			**PHOW**			**PHOW-RGB**		
Classe	NI	BOVW	SP	MP	BOVW	SP	MP	BOVW	SP	MP
Cannon	43	15.38	25.38	**26.92**	35.38	53.08	**56.15**	40.77	50	**51.54**
Octopus	35	28	**38**	36	28	**40**	**40**	30	**42**	38
Beaver	46	7.50	**23.13**	21.25	21.25	**36.25**	31.88	24.37	32.50	**36.88**
Ant	42	11.67	14.17	**15**	31.67	36.67	**40.83**	22.50	**33.33**	**33.33**
CougarBody	47	15.88	17.06	**19.41**	20	22.94	**24.12**	19.41	**35.29**	30.59
Crocodile	50	18.50	**31**	29	20	27	**31**	28	27	**28.50**
BGGoogle	467	18.12	19.06	**20.11**	18.81	20.18	**21.67**	18.44	20.55	**22.15**

Tabela A.1: Resultados obtidos pelos métodos BOVW, pirâmide espacial (SP) e método proposto (MP). Na tabela é mostrada a taxa de classificação correta média obtida pelos três métodos levando em consideração as classes com os melhores e piores resultados. Além disso, também é mostrado o número de imagens (NI) das classes apresentadas.

O gráfico da Figura A.1 mostra que utilizando o descritor local SIFT Denso, o método proposto obteve uma taxa de classificação correta média superior ao pirâmide espacial em 48% das classes, empatou em 12%, perdeu em 40%. Além disso, o MP foi superior ao método BOVW em 97% das classes, empatou em 1% e obteve um resultado inferior em 2% das classes. Na Figura A.2, são apresentados os resultados obtidos pelos métodos analisados, utilizando o descritor local PHOW. Com esse descritor, o método proposto e pirâmide espacial tiveram resultados muito semelhantes.

Também podemos perceber que, utilizando o descritor local SIFT Denso, o método proposto obteve uma taxa de classificação correta média superior a 90%, em 17 classes, enquanto os métodos pirâmide espacial e BOVW, em 16 e 7 classes, respectivamente. Com o descritor PHOW, o método proposto atingiu resultado superior a 90% em 23 classes, contra 22 e 12 classes dos métodos pirâmide espacial e BOVW, respectivamente. Dessa forma, podemos perceber a superioridade do método proposto em relação aos demais métodos analisados, independente do descritor local utilizado.

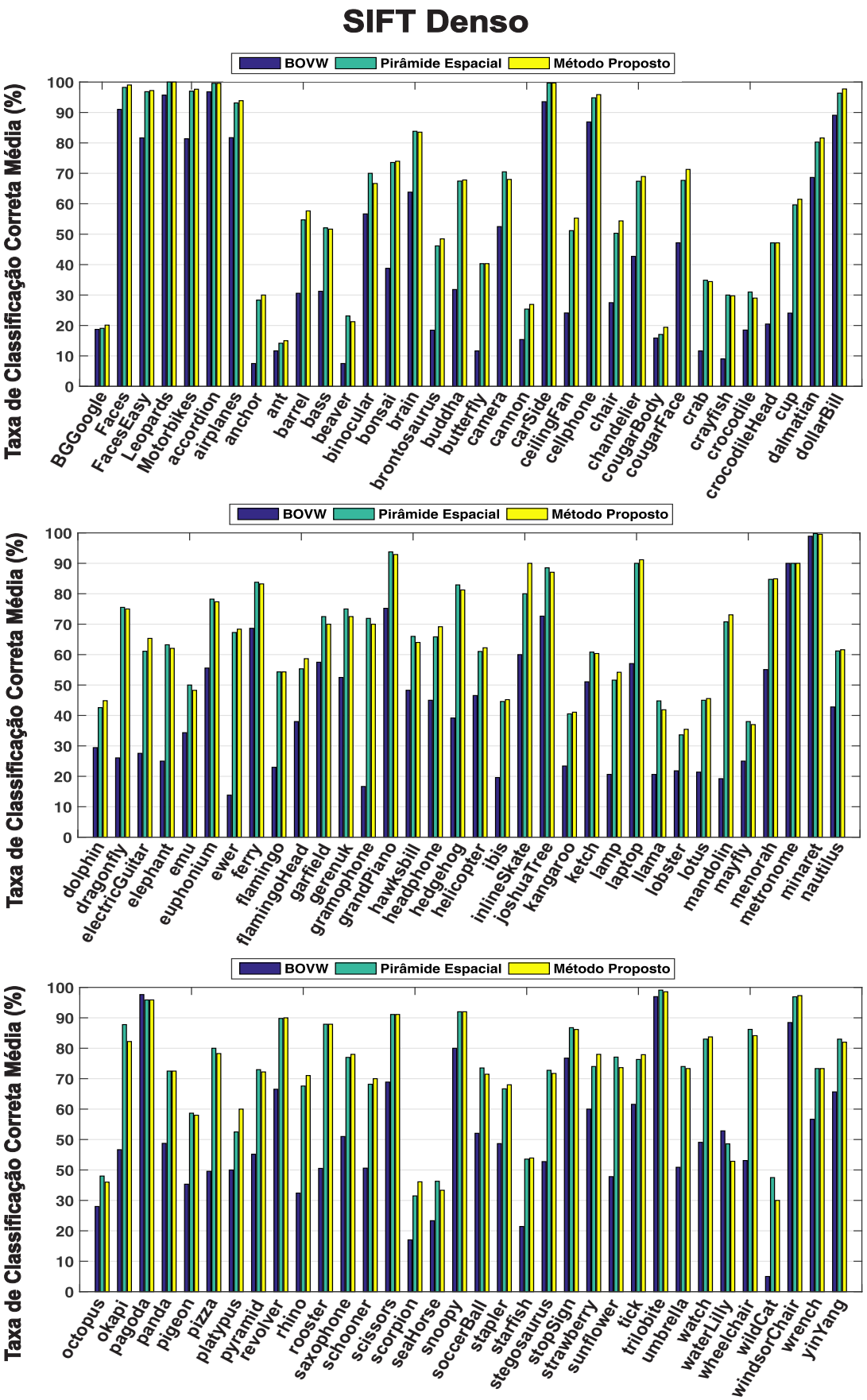


Figura A.1: Comparação de resultados entre os métodos BOVW, Pirâmide Espacial e método proposto, em relação a cada classe da base Caltech-101. A imagem mostra o melhor resultado de cada método como mostrado na Tabela 4.4, utilizando o SIFT Denso como descritor local.

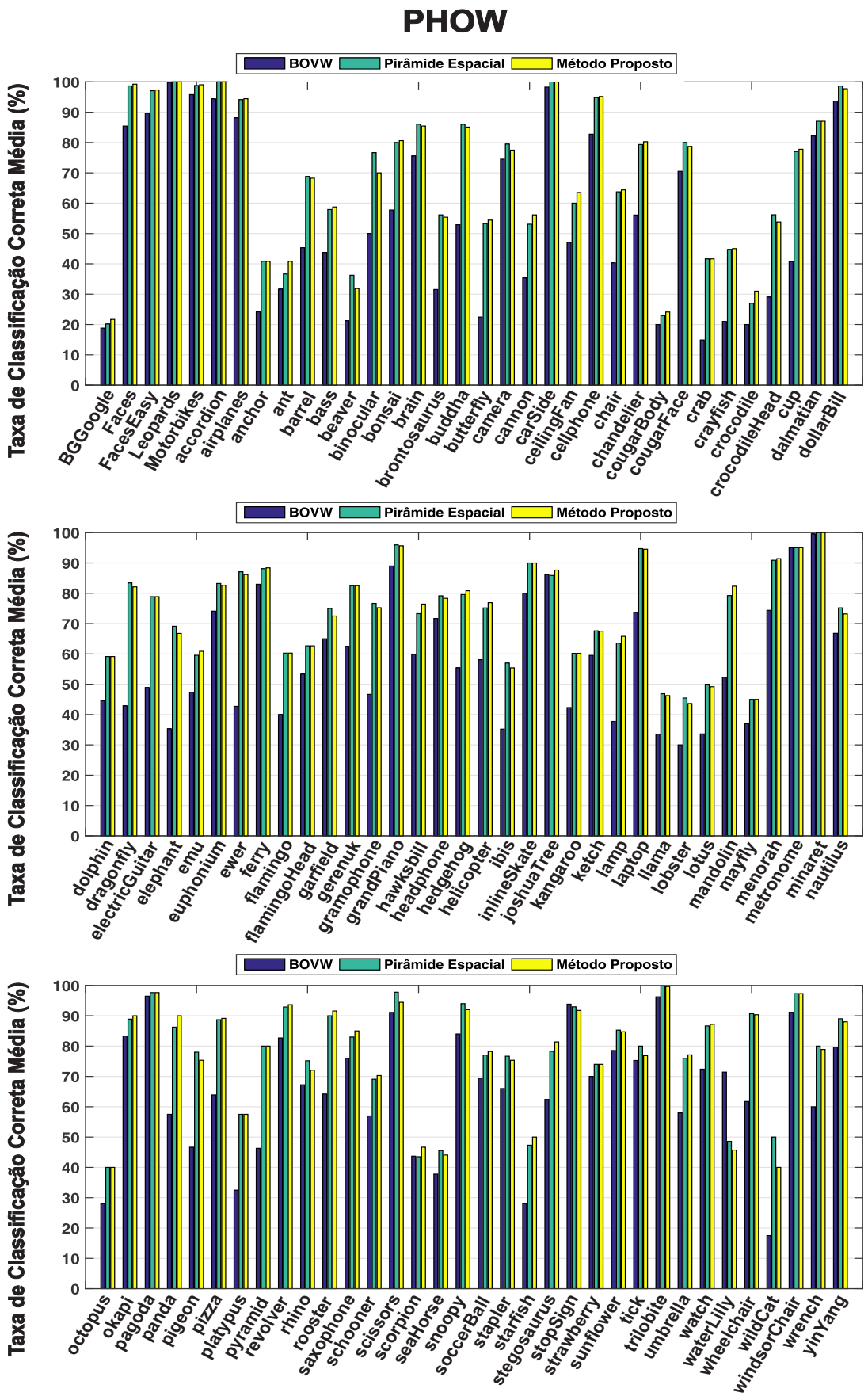


Figura A.2: Comparação de resultados entre os métodos BOVW, Pirâmide Espacial e método proposto, em relação a cada classe da base Caltech-101. A imagem mostra o melhor resultado de cada método como mostrado na Tabela 4.4, utilizando o PHOW como descritor local,.