UNIVERSIDADE ESTADUAL PAULISTA

Júlio de Mesquita Filho

Pós-Graduação em Ciência da Computação

Luis Claudio Sugi Afonso

Composição de Dicionários Visuais Utilizando Agrupamento de Dados por Floresta de Caminhos Ótimos

UNESP

2012

Luis Claudio Sugi Afonso

Prof. Dr. João Paulo Papa (Orientador)
Prof. Dr. Aparecido Nilceu Marana (Co-orientador)

Dissertação de Mestrado elaborada junto ao Programa de Pós-Graduação em Ciência da Computação - Área de Concentração em Sistemas de Computação como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

UNESP

2012

Departamento de Computação
Universidade Estadual Paulista

Luis Claudio Sugi Afonso

Fevereiro de 2013

Composição de Dicionários Visuais Utilizando Agrupamento de Dados por Floresta de Caminhos Ótimos

Banca Examinadora:

- Prof. Dr. João Paulo Papa (Orientador) João Paulo Papa
- Prof. Dr. Ricardo da Silva Torres (IC/Unicamp) - Titular
- Prof. Dr. Ivan Rizzo Guilherme (IGCE-UNESP) - Titular

Afonso, Luis Claudio Sugi.
Composição de dicionários visuais utilizando agrupamento de dados por flores de caminhos ótimos / Luis Claudio Sugi Afonso. - São José do Rio Preto: [s.n.], 2013.
57 f. : il. ; 30 cm.

Orientador: João Paulo Papa
Co-orientador: Marana, Aparecido Nilceu
Dissertação (mestrado) - Universidade Estadual Paulista, Instituto de Biociências, Letras e Ciências Exatas

1. Processamento de imagens - Técnicas digitais. 2. Reconhecimento de padrões. 3. Visão por computador. I. Papa, João Paulo. II. Nilceu, Marana Aparecido. III. Universidade Estadual Paulista, Instituto de Biociências, Letras e Ciências Exatas. IV. Título.

CDU -. 004.383.5

Ficha catalográfica elaborada pela Biblioteca do IBILCE
Campus de São José do Rio Preto - UNESP

.

Agradecimentos

Gostaria de agradecer meus pais *Claudio dos Santos Afonso* e *Sonia Ayaco Sugi Afonso* por todo o apoio que me deram não apenas durante o mestrado, mas durante todo o percurso percorrido até o momento. Obrigado!!!

Agradeço aos meus orientadores *Profs. João Paulo Papa* e *Aparecido Nilceu Marana* primeiramente por terem aceitado serem meus orientadores durante essa passagem pela UNESP, tanto na graduação, quanto na pós-graduação. Também agradeço pela ajuda e apoio durante essas etapas, durante o desenvolvimento dos projetos e, é claro, por terem aceitado e apoiado minha participação no intercâmbio no Canadá.

Também agradeço a UNESP por possibilitar a realização do mestrado e por conceder a bolsa durante o curso.

Agradeço à *Profa. Svetlana Yanushkevich* da Universidade de Calgary-Canadá por ter aceitado ser minha orientadora e por ter me auxiliado antes e durante o intercâmbio em Calgary. Também agradeço a todos do *Biometric Technologies Laboratories* por terem me recebido bem e pela amizade durante o período que trabalhei por lá.

Claro que não posso me esquecer da minha namorada, *Jéssica Akemi Matsuoka*, por ter estado comigo em todos os momentos e pela felicidade que me trouxe. Tem sido muito bom ter você ao meu lado. Obrigado por tudo! Te amo!

Ao pessoal do LCAD/RECOGNA, *Clayton, Mizobe, Cristiano, Alan Peixinho, Luis Augusto, Alex, Adriana,...*, obrigado! Obrigado aos meus amigos do CTI, *Bruno Pereira Matheus, Wallace Sato*, e todos que fizeram parte do grupo do CTI.

Também agradeço a todos que me ajudaram direta ou indiretamente..

Desculpe se esqueci de mencionar alguém.

Enfim, obrigado a todos que fizeram parte de mais essa etapa!!!!

Resumo

Categorização de imagens utilizando Dicionário de Palavras Visuais tem recebido grande atenção pelas comunidades de visão computacional e processamento de imagens. Nesta abordagem, cada imagem é representada por um conjunto de pontos invariantes, os quais são mapeados no espaço de Hilbert, o qual é uma extensão do plano Euclideano e espaço 3D tendo qual quer número finito ou infinito de dimensões, representando um dicionário visual composto das características mais representativas de um conjunto de imagens. Contudo, o principal problema de tal abordagem é encontrar um dicionário que seja compacto e, ao mesmo tempo, representativo. Encontrar tal dicionário de maneira automática, sem auxílio de um usuário, é uma tarefa ainda mais difícil. Neste trabalho, é proposto um método para encontrar o dicionário de maneira automática empregando um algoritmo baseado em grafos denominado Floresta de Caminhos Ótimos, o qual não necessita da dimensão do dicionário para encontrá-lo. Os experimentos envolveram o uso de 3 bases de imagens de objetos variados e realizando-se uma comparação entre a técnica apresentada e as técnicas K-médias e Seleção Aleatória. A comparação avaliou o tempo necessário para que cada técnica compute os dicionários e a taxa de acerto proporcionada pelos dicionários. Os resultados experimentais monstraram que o algoritmo Floresta de Caminhos Ótimos é uma alternativa a ser empregada na técnica Dicionário de Palavras Visuais, uma vez que as taxas de acerto são similares às demais técnicas, possui vantagem quando dicionários de alta dimensão devem ser calculados e, principalmente, não necessita que a dimensão do dicionário visual seja definido *a priori*.

Palavras-chave: Floresta de Caminhos Ótimos, Agrupamento de Dados, Dicionário Visual.

Abstract

Image categorization by means of bag of visual words has received increasing attention by the image processing and vision communities in the last years. In these approaches, each image is represented by invariant points of interest which are mapped to a Hilbert Space, which is an extension of traditionals Euclidean plane and 3D space having any finite or infinite number of dimensions, representing a visual dictionary which aims at comprising the most discriminative features in a set of images. Notwithstanding, the main problem of such approaches is to find a compact and representative dictionary. Finding such representative dictionary automatically with no user intervention is an even more difficult task. In this work, we propose a method to automatically find such dictionary by employing a recent developed graph-based não-supervisionado algorithm called Optimum-Path Forest, which does not make any assumption about the visual dictionary's size. Experiments were performed on 3 different databases of different objects in order to compare OPF não-supervisionado, K-means and Random Selection. The comparison assessed the time for each technique to compute the visual dictionaries and the accuracy rate when such visual dictionaries are used. The experimental results showed that OPF não-supervisionado is an alternate algorithm for the visual dictionary generation, since accuracy rates are similar, presents a time advantage when high-dimension dictionaries have to be computed and does not require visual dictionary dimension prior its computing.

Keywords: Optimum-Path Forest, Data não-supervisionado, Visual Dictionary.

Sumário

1 Introdução

Com a popularização das câmeras em dispositivos eletrônicos e o surgimento de ferramentas para compartilhamento de fotos fez com que o número de imagens disponíveis tivesse um aumento considerável [Garg, Vempati e Jawahar 2011]. Com esse crescimento, há também uma maior variedade de imagens de objetos ou cenários, por exemplo. Consequentemente, a recuperação de imagens torna-se um processo lento e, muitas vezes, retornando imagens não relacionadas a busca original. A lentidão na etapa de recuperação deve-se ao número de comparações a serem realizadas para se encontrar as imagens mais semelhantes à imagem buscada. Além disso, tem-se o tempo necessário para calcular o nível de semelhança entre as imagens, o qual depende de como as imagens são representadas e da função de semelhança empregada.

Para contornar esse tipo de problema, podem-se utilizar representações mais simples para cada imagem por meio de vetores de características de baixa dimensão, por exemplo. Vetores de características de baixa dimensão proporcionam um ganho de tempo, porém, podem ocasionar em uma perda na taxa de acerto por não possuírem informações suficientes para discriminar as imagens em classes, tornando o sistema ineficiente.

Contudo, há técnicas simples e ao mesmo tempo robustas o suficiente para gerar uma representação simples e manter o reconhecimento eficiente. Uma delas é o "Dicionário de Palavras Visuais"(*Bag-of-Visual Words - BoVW*) [Csurka et al. 2004], a qual tem sido aplicada para realizar a categorização de imagens. Essa técnica mostra-se simples, pois cria uma representação para cada imagem utilizando um grupo de descritores obtidos a partir de um conjunto de imagens de referência e, ao mesmo tempo, é robusta, uma vez que a representação possui uma certa tolerância a transformações, como escala e orientação.

O Dicionário de Palavras Visuais gera o vetor de características por meio de 3 etapas: (i) descritores são extraídos de um conjunto de imagens de referência (conjunto de treinamento), (ii) realiza a seleção dos descritores escolhendo os mais discriminantes e descartando redundâncias e ruídos, e (iii) realiza a quantização de cada imagem por meio do conjunto de descritores selecionados (dicionário visual) para gerar o vetor de características de cada imagem. Os vetores de características são aplicados em classificadores para realizar seu treinamento e identificar novas imagens.

Outro ponto a ser considerado é que cada imagem possui um número diferente de descritores. Consequentemente, a dimensão dos vetores de características será diferente, dificultando a utilização de classificadores de padrões e tornando a classificação um processo custoso. O alto custo deve-se ao fato da necessidade da utilização de alguma técnica de *matching*, a qual deverá realizar o casamento de cada descritor. Aplicando BoVW, é possível gerar uma representação de mesma dimensão para cada imagem permitindo, então, a utilização de classificadores.

O dicionário visual é constituído de palavras visuais, as quais consistem em descritores extraídos de regiões consideradas estáveis com relação à escala e rotação. As palavras visuais são selecionadas utilizando-se um algoritmo de agrupamento, uma vez que nem todas podem ser utilizadas devido a restrições de armazenamento e redundância. A seleção faz-se necessária, pois o dicionário visual deve ser o mais discriminativo possível para se obter resultados satisfatórios durante a categorização das imagens.

Como mencionado, algoritmos de seleção ou de agrupamento podem ser empregados para realizar a seleção das palavras visuais. Alguns deles desempenham a seleção de maneira bem simples, como selecionando aleatoriamente as palavras visuais. Contudo, tal critério pode descartar palavras visuais que seriam mais importantes ao dicionário. Uma alternativa seria a utilização de algoritmos de agrupamento, os quais calculam a palavra visual que melhor representa cada aglomerado de palavras visuais, em que K-médias é a técnica mais comum para tal finalidade [Bosch, Zisserman e Muñoz 2008, Liu e Shah 2007, Marszałek et al. 2007, Sande, Gevers e Snoek 2010, Yang et al. 2007].

Alguns algoritmos de agrupamento requerem o número de aglomerados antes do cálculo dos pontos médios de cada um deles ser realizado. Neste projeto, o número de

aglomerados representa o número de palavras visuais a serem utilizadas no dicionário visual. A escolha desse número é importante, pois ele tem impacto tanto na precisão do sistema de categorização, quanto no tempo necessário para encontrar a qual classe uma determinada imagem de entrada pertence. Por isso, a utilização de um algoritmo desse tipo é custoso quando se necessita de um sistema preciso, já que é necessário encontrar a dimensão do dicionário que forneça os melhores resultados.

Uma alternativa para o problema de seleção da dimensão do dicionário visual é a utilização de algoritmos de agrupamento que não necessitem do número de aglomerados como parâmetro inicial. Além de não necessitar da dimensão do dicionário visual antes de realizar seu cálculo, tais algoritmos podem realizar essa tarefa tão rapidamente quanto um algoritmo que necessita desse valor. Além disso, o dicionário resultante pode apresentar menor dimensão e melhores resultados quando comparados com as demais técnicas.

Recentemente, um novo algoritmo de classificação baseado em grafos denominado *Optimum-Path Forest* (OPF) [Papa, Falcão e Suzuki 2009] foi apresentado, possuindo suas versões supervisionada e não-supervisionada. Em ambas as versões, cada amostra é representada como sendo um nó de um grafo completo e, cujas arestas são ponderadas pela distância entre os vetores de características de cada amostra. Durante a etapa de treinamento, é realizada uma competição entre os nós mais representativos, denominados *protótipos*. Os protótipos disputam cada amostra do conjunto, oferecendo custos ótimos e conquistando-os, caso seu custo seja o melhor entre todos os oferecidos. Aquele que oferecer o melhor custo, associará seu rótulo e custo à amostra conquistada. Ao final desse processo, obtém-se uma floresta de caminhos ótimos, em que cada árvore é enraizada por um protótipo.

A maneira como os protótipos são definidos é distinta para cada versão. Na versão supervisionada, os protótipos são definidos como sendo os nós de classes distintas mais próximos, de acordo com o valor de suas arestas. Devido a esse critério, uma classe pode possuir mais de uma árvore representando-a. Na versão não-supervisionada, faz-se uso de uma função de densidade probabilidade (fdp), onde os protótipos serão as amostras com os maiores valores. A partir dos valores de fdp e dos protótipos, são computados agrupamentos, os quais serão as árvores de caminhos ótimos. Diferentemente da versão

supervisionada, cada classe é representada por uma única árvore.

Por meio do OPF não-supervisionado, é possível encontrar agrupamentos em conjuntos de dados sem a necessidade de definir seu número *a priori*, além de encontrar os elementos mais representativos. Essa característica apresenta vantagens quando aplicadas em técnicas como BoVW, uma vez que o OPF não-supervisionado pode encontrar melhores agrupamentos e em menor tempo quando comparado à outras técnicas de agrupamento.

O objetivo deste trabalho é empregar a versão não-supervisionada na etapa de cálculo de dicionários visuais e avaliar seu desempenho, tanto com relação a precisão proporcionada ao sistema, quanto com relação ao tempo necessário para calcular cada dicionário visual, comparando com os demais algoritmos empregados.

O desempenho da construção de dicionários visuais por OPF não-supervisionado será comparado com outras técnicas de criação de dicionários (seleção aleatória e K-médias) por meio da taxa de classificação do classificador Bayesiano.

Esse trabalho está organizado da seguinte maneira: No Capítulo 2, são descritas as duas técnicas empregadas para realizar a extração dos descritores das imagens. No Capítulo 3, é apresentada e descrita cada uma das etapas que constituem a técnica de categorização de imagens Dicionário de Palavras Visuais. O Capítulo 4 apresenta a técnica de agrupamento não-supervisionado OPF não-supervisionado. As etapas para a realização dos estudos experimentais e seus resultados são apresentados no Capítulo 6, enquanto que no Capítulo 7, são apresentadas as conclusões dos resultados obtidos. Finalmente, no Capítulo 8, são descritas as atividades desenvolvidas durante o mestrado.

2 Descritores de Imagens

Este capítulo introduz as técnicas empregadas neste trabalho para realizar a extração de descritores. SIFT [Lowe 2004] foi escolhida por ser a mais comum entre os trabalhos que utilizam BoVW. Já a técnica SURF foi selecionada por ser uma alternativa ao SIFT e pelos resultados apresentados pelos seus autores [Bay et al. 2008] .

2.1 Transformada de Características Invariantes à Escala

A Transformada de Características Invariantes à Escala (*SIFT*) apresentada por Lowe [Lowe 2004] recebeu esse nome por transformar dados de uma imagem em coordenadas invariantes à escala relacionadas às características locais, e tem como objetivo extrair características de imagens que possibilitem a recuperação de cenas ou objetos eficientemente. Para isso, são extraídas características invariantes a escala e rotação da imagem, e que também sejam estáveis com relação a ruídos.

O processo de obtenção das características utilizando SIFT pode ser dividido em quatro etapas apresentadas no diagrama da Figura 2.1 e descritas a seguir.

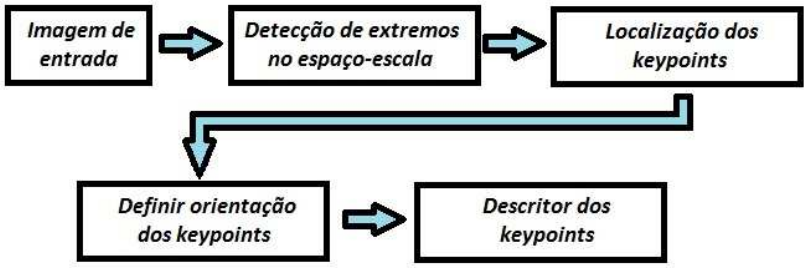


Figura 2.1: Etapas de desenvolvimento do SIFT.

2.1.1 Detecção de Extremos no Espaço-escala

Nesse estágio, são identificadas as localizações e escalas de potenciais pontos importantes (*keypoints*) que sejam invariantes a escala e orientação. A identificação de tais localizações é feita a partir da busca por características que sejam estáveis utilizando uma função contínua de escala ou espaço-escala [Lowe 2004].

O espaço-escala de uma imagem é definido como *L(x, y, σ)*, o qual é produzido a partir da convolução de uma função Gaussiana com escala variável, *G(x, y, σ)*, tendo como entrada uma imagem *I(x, y)*:

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

em que $\otimes$ é o operador de convolução aplicado ao *pixel* $p(x, y)$, e

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma} e^{-\frac{(x^2+y^2)}{2\sigma^2}}, \tag{2.2}$$

em que a variável σ representa a escala, ou seja, desvio padrão da função Gaussiana. Por isso, a atribuição de diferentes valores a σ possibilita a criação de imagens com diferentes níveis de suavização. Cada imagem suavizada representará um espaço-escala diferente, o qual é utilizado nas etapas de detecção de pontos que sejam estáveis na imagem.

Os pontos considerados estáveis são os pontos extremos do espaço-escala da diferença da função Gaussiana, *D(x, y, σ)*, aplicada em I. A diferença da função Gaussiana pode ser calculada a partir da diferença de duas escalas próximas separadas por um fator multiplicativo constante k:

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

Dada uma imagem *I(x, y)*, são aplicadas a ela funções do tipo Gaussiana com diferentes valores de σ. Cada função irá gerar uma imagem com diferentes níveis de suavização. As

imagens suavizadas são subtraídas umas das outras obtendo-se, então, a imagem com Diferença de Gaussianas (*Difference of Gaussians - DoG*).

Após esse processo, a imagem I é reduzida e novamente são aplicadas as funções Gaussianas para se obter as imagens DoG. Cada conjunto de imagens Gaussianas e imagens DoG é denominado *oitavo*. Cada oitavo é referente a uma imagem original redimensionada, conforme ilustra a Figura 2.2.

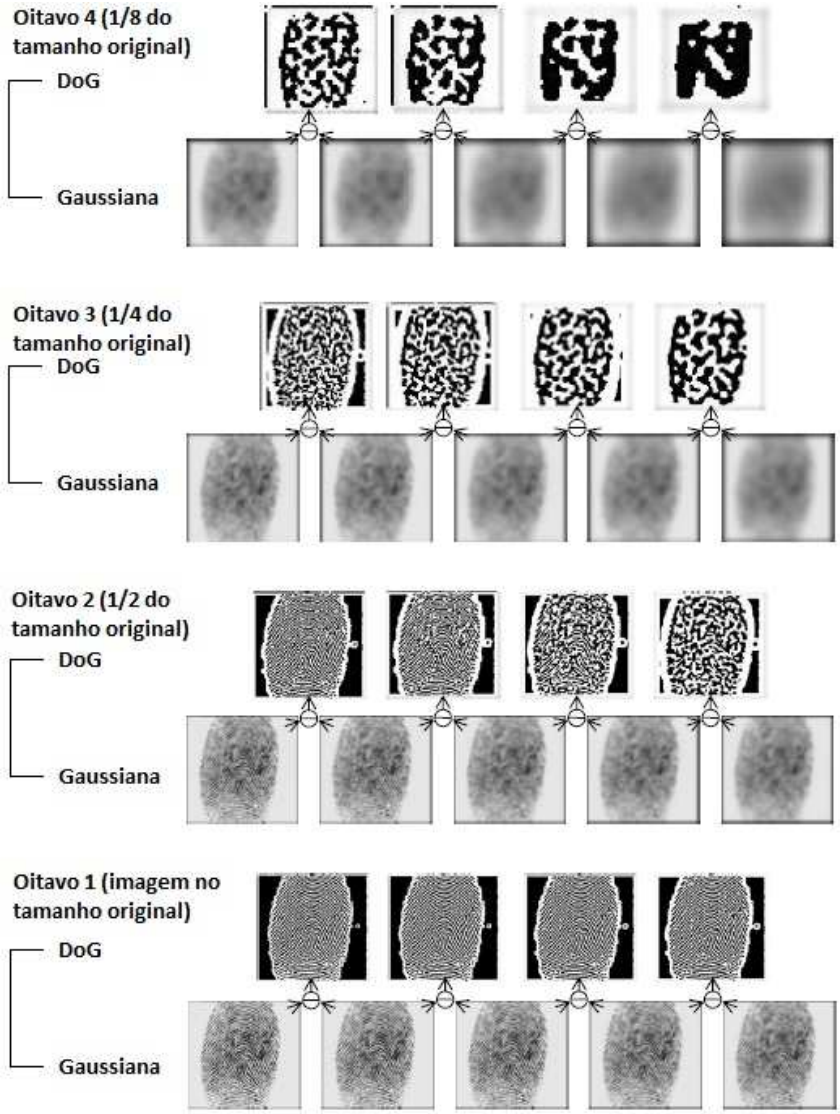


Figura 2.2: Exemplo do processo de obtenção das imagens DoG. Extraído e adaptado de Pankanti e Jain [Park e Jain 2008].

Uma vez obtidas todas as imagens DoG de todos os oitavos, inicia-se a busca pelos extremos locais, os quais podem ser máximos ou mínimos locais. Para definir os extremos locais, uma região de tamanho 3×3 é delimitada na imagem. O ponto central é comparado com todos os seus vizinhos e com os pontos das escalas acima e abaixo da imagem que está sendo trabalhada, conforme é visto na Figura 2.3. O processo é realizado para todas as escalas de cada oitavo, exceto a primeira e a última pelo fato de não haver pontos suficientes para realizar a comparação. Caso o ponto seja máximo ou mínimo, ele é selecionado como candidato a ser um *keypoint*. Caso contrário, o ponto é descartado [Lowe 2004].

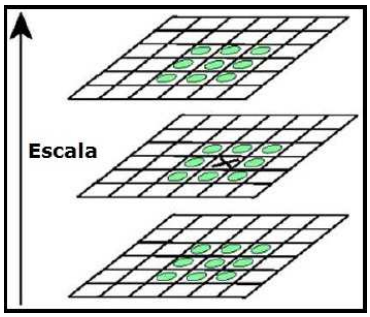


Figura 2.3: Processo de comparação para encontrar os extremos das imagens. Extraído de Lowe [Lowe 2004].

2.1.2 Localização dos *Keypoints*

Em sua implementação inicial, Lowe [Lowe 2004] simplesmente utilizava a localização e a escala do ponto central da amostra para localizar os *keypoints*. Para melhorar o reconhecimento e a estabilidade das características utilizadas, foi aplicada a expansão de Taylor, ilustrada na Figura 2.4 na função espaço-escala, *D(x, y, σ)*, utilizando como origem o ponto de amostra (ponto selecionado):

$$D(x) = D + \frac{\partial D^T}{\partial x}x + \frac{1}{2}x^T\frac{\partial^2 D}{\partial x^2}x, \tag{2.5}$$

em que D e suas derivadas são avaliadas no ponto de amostra e $\mathbf{x}=(x, y, \sigma)^T$ é a compensação. A localização do extremo $\hat{x}$ é determinada a partir do uso da derivada da função e igualando-a zero [Lowe 2004]:

$$\hat{x} = -\frac{\partial^2 D}{\partial x^2}^{-1} \frac{\partial D}{\partial x}. \tag{2.6}$$

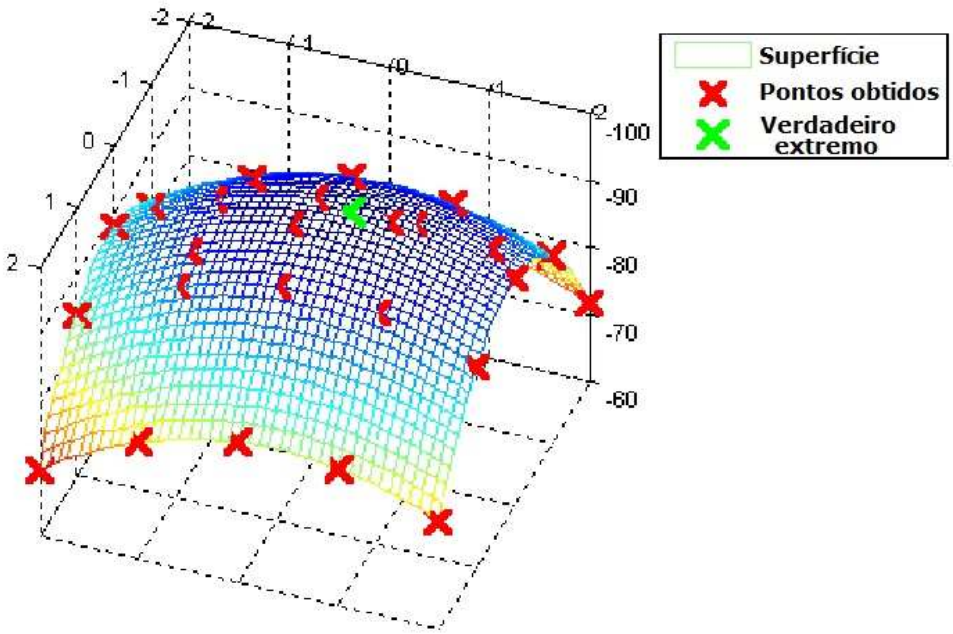


Figura 2.4: Exemplo de interpolação realizada utilizando os candidatos a *keypoints*. Adaptado de Sinha [Sinha 2010].

Nem todos os candidatos a *keypoint* detectados podem ser utilizados, pois alguns possuem baixo contraste e sua localização na borda pode fazer com o *keypoint* sofra grande influência de ruídos. Para definir quais candidatos são realmente estáveis, utiliza-se uma matriz Hessiana H de tamanho 2×2, a qual é calculada na localização e escala do *keypoint*. As derivadas são estimadas pela diferença de pontos vizinhos.

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

A partir dos auto-valores de H é possível eliminar os pontos suscetíveis a ruídos. Sendo, α e β os auto-valores de maior e menor magnitude, respectivamente, e:

$$Tr(H) = D_{xx} + D_{yy} = \alpha + \beta,$$
$$Det(H) = D_{xx}D_{yy} + (D_{xy})^2 = \alpha\beta,$$

em que Tr é o *trace* e Det o determinante. Nesse caso, r denota a proporção entre o maior e o menor auto-valor, ou seja, $\alpha = r\beta$. Assim, temos que:

$$\frac{Tr(H)^2}{Det(H)} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r}.$$

Portanto, a verificação da proporção entre as principais curvaturas é dada por

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

Com isso, os *keypoints* que possuírem uma taxa de proporção entre os auto-valores maior do que um determinado limiar são descartados.

2.1.3 Orientação

Ao descritor dos *keypoints* é associada sua orientação baseada nas propriedades locais da imagem com o intuito de proporcionar invariância com relação a rotação da imagem. Dada uma imagem suavizada S, para cada amostra *S(x, y)* são calculadas a magnitude do gradiente e orientação a partir da diferença de *pixels*, conforme as Equações 2.7 e 2.8, respectivamente:

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \tag{2.7}$$

$$\theta(x,y) = \arctan\left(\frac{(L(x,y+1) - L(x,y-1))}{(L(x+1,y) - L(x-1,y))}\right). \tag{2.8}$$

Cada *keypoint* possuirá seu respectivo histograma de orientação referente ao gradiente de orientações das amostras vizinhas. A vizinhança é definida por uma janela circular com

peso definido por uma função Gaussiana (Figura 2.5). Para cada *pixel*, é atribuído um peso definido pela sua magnitude do gradiente e pelo peso da função Gaussiana citada.

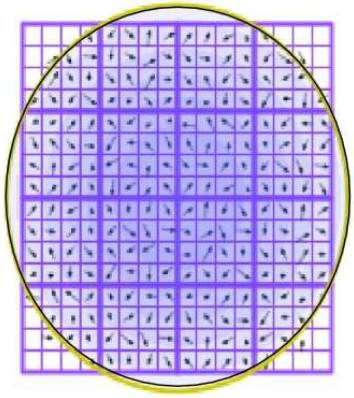


Figura 2.5: Região delimitada por uma função de peso Gaussiano, círculo amarelo. Extraído de Clemons [Clemons 2012].

Os picos no histograma de orientação indicam direções dominantes dos gradientes locais. Os picos com até 80% do valor do maior pico detectado no histograma são utilizados para criar novos *keypoints*. Por isso, em locais onde há vários picos, existirão vários *keypoints* com a mesma localização e escala, porém com diferentes orientações [Lowe 2004]. A Figura 2.6 apresenta um exemplo de histograma de orientação. Neste histograma, são detectados dois picos, sendo um no intervalo de orientações entre 20° e 29° e outro no intervalo entre 300° e 309°. O segundo pico também é utilizado para caracterizar o *keypoint* por possuir um valor maior do que 80% do pico principal.

2.1.4 Descritor dos *Keypoints*

O cálculo do descritor dos *keypoints* inicia-se com o cálculo da magnitude do gradiente e orientação das amostras vizinhas ao *keypoint*. Para que haja invariância à rotação, as coordenadas do descritor e as orientações dos gradientes são rotacionados em relação à orientação do *keypoint*.

Uma função de peso Gaussiana é utilizada para atribuir pesos às magnitudes dos gradientes das amostras vizinhas do *keypoint*. A função Gaussiana é, então, utilizada para evitar mudanças repentinas no descritor e diminuir o efeito de gradientes que estão

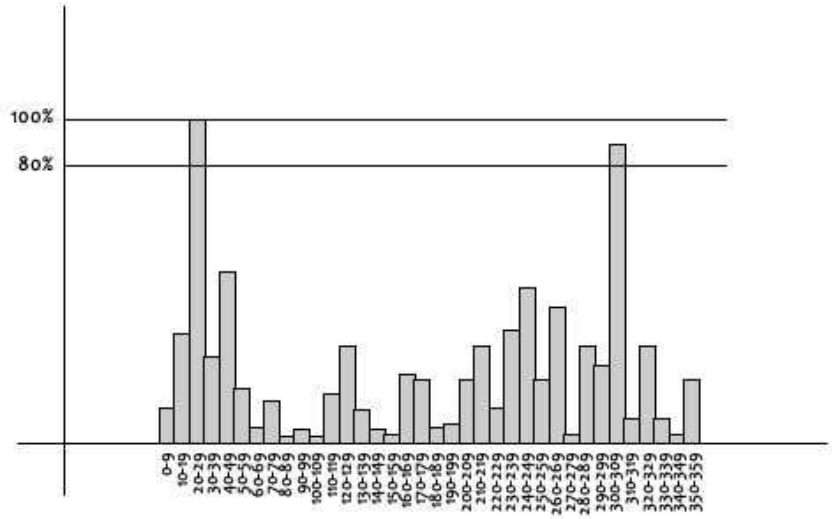


Figura 2.6: Exemplo de histograma obtido. Extraído de Sinha [Sinha 2010].

mais afastados do centro da janela. O quadro do lado direito na Figura 2.7 apresenta o descritor do *keypoint*, o qual é definido pelo histograma de orientações e pela magnitude dos gradientes de cada uma das 16 janelas.

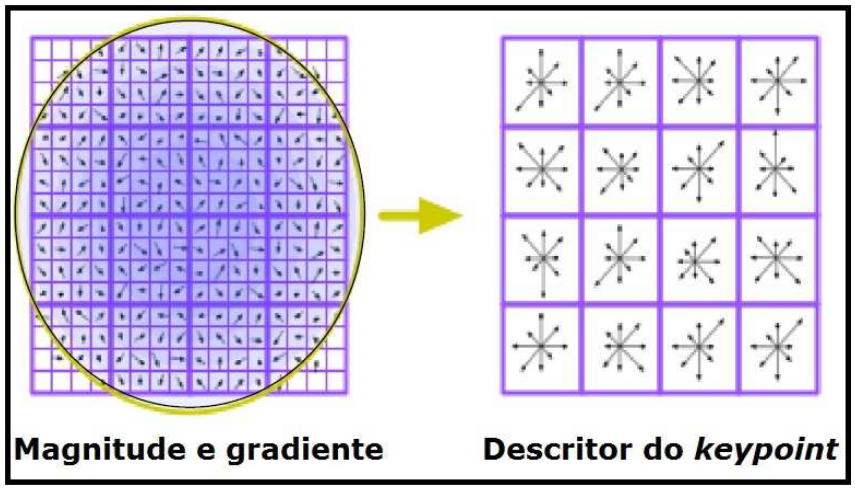


Figura 2.7: Descritor SIFT. Extraído de Clemons [Clemons 2012].

Cada uma das 16 janelas possuem 8 características referentes a cada uma das 8 divisões de 45° definidas para realizar a construção do histograma. Com isso, tem-se o descritor formado por 128 elementos, os quais são normalizados para uma escala unitária.

2.2 *Speeded-Up Robust Features*

Proposta por Bay, Tuytelaars e Van Gool [Bay et al. 2008], a técnica possui próposito semelhante à técnica apresentada por Lowe [Lowe 2004], uma vez que ambas buscam extrair pontos de interesse que sejam invariantes à escala e rotação. Os autores buscam desenvolver uma técnica que possua um método simples para realizar a detecção dos pontos de interesse, mas que seja ao mesmo tempo preciso e, também, que o descritor final possua baixa dimensão e que seja suficientemente discriminativo.

Essa técnica pode ser dividida em duas etapas: detecção dos pontos de interesse e cálculo dos descritores, descritas a seguir.

2.2.1 Detecção de Pontos de Interesse

A detecção de pontos de interesse é realizada por meio de uma aproximação da matriz Hessiana aplicada na imagem integral da imagem de entrada, com o intuito de reduzir o tempo computacional.

A imagem integral na localização (x,y) é igual a soma de todos os pixels que estão acima e à esquerda de (x,y) incluindo o próprio:

$$ii(x,y) = \sum_{x' \leq x, y' \leq y} = i(x', y'), \tag{2.9}$$

em que (ii) é a imagem integral do *pixel* localizado em (x,y) e $(x',\ y')$ representam as coordenadas de todos os *pixels* que estão acima e à esquerda de (x,y) na imagem original i. A Figura 2.8 apresenta como é realizado o somatório dos *pixels* pertencentes à região em azul.

A matriz Hessiana é então responsável por detectar estruturas na forma de bolhas cujo determinante é máximo. Também por meio da matriz Hessiana, define-se a escala da imagem na qual a estrutura foi detectada. Para isso são utilizados filtros na forma de caixa de dimensões variadas para calcular as respostas de cada região da imagem em diferentes escalas.

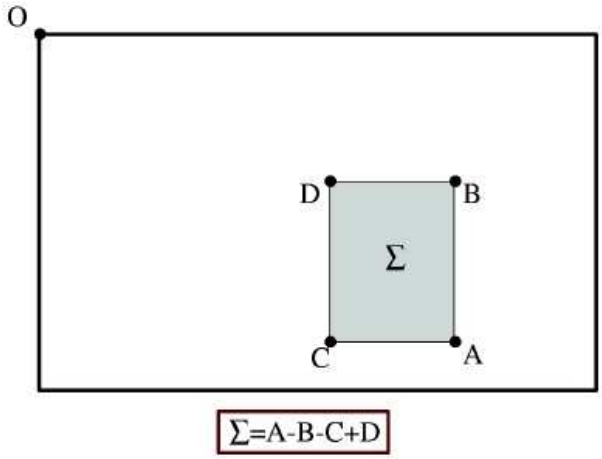
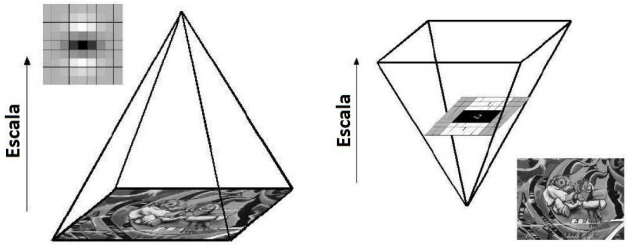


Figura 2.8: O somatório dos valores dos *pixels* dentro da região em azul é definido pela soma e subtração dos valores dos 4 *pixels* que definem a região. Extraído de [Bay et al. 2008].

Espaços-escala são normalmente implementados na forma de uma pirâmide de imagens, em que as imagens são suavizadas por um filtro Gaussiano e, posteriormente, reduzidas para se atingir os níveis mais altos da pirâmide, ou seja, um mesmo filtro é aplicado nas diferentes escalas de uma dada imagem de entrada. Cada nível do espaço-escala é denominado oitavo e representam uma série de mapas de respostas de uma imagem de entrada para um filtro de tamanho crescente. Cada oitavo é sub-dividido em um número constante de níveis de escala.

Diferentemente da abordagem convencional, a técnica SURF realiza a análise por meio do redimensionamento dos filtros e aplicando-os diretamente na imagem de entrada. As Figuras 2.9(a) e (b) apresentam a diferença entre a abordagem adotada por Lowe e a abordagem adotada por Bay, Tuytelaars e Van Gool, respectivamente.

Assim como Lowe [Lowe 2004], Bay, Tuytelaars e Van Gool utilizam uma aproximação do filtro Gaussiano, porém, na forma de caixas. O tamanho inicial do filtro é uma janela de tamanho 9×9 representando uma escala $s = 1{,}2$ (correspondente a um $\sigma = 1{,}2$) [Bay et al. 2008], a qual representa a menor das escalas empregadas. No primeiro oitavo do espaço-escala, o filtro inicial tem suas dimensões aumentadas em 6, ou seja, são aplicados filtros de dimensões 9×9, 15×15, 21×21 e 27×27. A cada novo oitavo, o aumento no filtro dobra, ou seja, o primeiro oitavo tem um redimensionamento em 6, o



(a) (b)

Figura 2.9: Diferença entre as abordagens adotadas para as técnicas: (a) SIFT e (b) SURF. Extraídos de [Bay et al. 2008].

segundo em 12, o terceiro em 24, e assim sucessivamente.

Para localizar os pontos de interesse na imagem e com relação à escala, é aplicada uma supressão não-máxima em uma vizinhança 3×3×3. Os máximos do determinante da matriz Hessiana são interpolados em escala e imagem-espaço utilizando o método proposto por Brown e Lowe [Brown e Lowe 2002].

2.2.2 Cálculo dos Descritores

O descritor calculado codifica a distribuição do conteúdo de intensidade da vizinhança do ponto de interesse. Para isso, são calculadas as respostas nas direções x e y a partir da distribuição de primeira ordem da *wavelet* de Haar gerando um descritor de 64 dimensões. Segundo o autor, isso reduz o tempo para cálculo dos descritores e comparação dos mesmos, além de aumentar a robustez.

Para se obter o descritor, é necessário definir a orientação do ponto de interesse baseada nas informações de uma região circular de raio $6s$ em torno do ponto, em que s é a escala na qual o ponto de interesse foi encontrado. Uma vez definida a região, realiza-se o cálculo das respostas nas direções x e y utilizando os filtros apresentados na Figura 2.10. Após o cálculo, as respostas são ponderadas por uma Gaussiana ($\sigma = 2s$) centralizada no ponto de interesse e representada como pontos em um espaço onde as respostas horizontais são representadas no eixo das abscissas e as respostas verticais no eixo das ordenadas.

A orientação dominante é estimada por meio do cálculo da soma de todas as repostas dentro de uma janela deslizante de dimensão $\frac{\pi}{3}$. A soma das respostas horizontais e

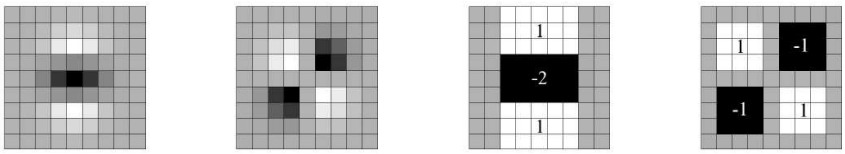


Figura 2.10: Filtros: os dois filtros da esquerda representam a derivada parcial de ordem 2 da função Gaussiana, enquanto que os dois filtros da direita representam suas respectivas aproximações. Extraído de [Bay et al. 2008].

verticais resultará em um vetor de orientação local. O maior entre todos os vetores definirá a orientação do ponto de interesse.

O descritor é então calculado para uma região delimitada por um quadrado de dimensão $20s$, centralizado no ponto de interesse e orientado pela orientação do ponto de interesse. Essa região é dividida em sub-regiões de dimensão 4×4, em que para cada sub-região são calculadas as respostas para a *wavelet* de Haar nas direções horizontal e vertical com relação a orientação do ponto de interesse, definida na etapa anterior. As respostas horizontal e vertical são representadas como d_x e d_y, respectivamente.

Para aumentar a robustez com relação a transformações geométricas, as respostas d_x e d_y são primeiramente ponderadas por uma Gaussiana ($\sigma = 3.3s$) centralizada no ponto de interesse. O descritor também fornece informações sobre a polaridade ao calcular a soma dos valores absolutos de d_x e d_y, resultando em $|d_x|$ e $|d_y|$. Com isso, cada sub-região possuirá um descritor de dimensão 4, resultando em um descritor de dimensão 64.

2.3 Considerações Finais

O capítulo apresentou as técnicas de extração de descritores SIFT e SURF, assim como uma explanação de cada uma de suas etapas. SIFT e SURF realizam a análise das imagens em diferentes escalas aplicando filtros do tipo Gaussiana para identificar regiões estáveis a transformações de escala e orientação. Os filtros são redimensionados a cada iteração, simulando diferentes escalas. Para a técnica SIFT, orientação e gradiente dos vizinhos do ponto de interesse são utilizados para realizar o cálculo do descritor final de 128 dimensões. Para a técnica SURF, são utilizadas a orientação e as respostas para a *wavelet* de Haar,

resultando em um descritor de 64 dimensões. No capítulo seguinte, são apresentados os algoritmos de agrupamento estudados e avaliados durante os experimentos.

3 Dicionário de Palavras Visuais

Este capítulo apresenta a técnica Dicionário de Palavras Visuais e cada uma das etapas que constituem tal técnica.

3.1 Visão Geral

A técnica de *Dicionário de Palavras Visuais (Bag-of-Visual Words - BoVW)* [Csurka et al. 2004] foi proposta para realizar a categorização de imagens e baseia-se em Dicionário de Palavras (*Bag-of-Words - BoW*), técnica para categorização de textos, possuindo etapas semelhantes, porém, empregando algoritmos próprios para a área na qual são aplicadas.

Devido ao crescimento do número de textos, técnicas de categorização têm sido propostas com o intuito de auxiliar a recuperação dos mesmos. Dentre elas estão aquelas baseadas nas Máquinas de Vetores de Suporte (*Support Vector Machines - SVM*) [Joachims 1998, Tong e Koller 2002] e em BoW. As técnicas baseadas em SVM tratam o problema de categorização de textos como sendo linearmente separável, em que as classes são os assuntos a que ada texto pertence.

BoW é uma técnica simples e rápida, a qual realiza o cálculo de um vetor de características para cada texto, baseando-se na frequência de cada palavra de um conjunto denominado *vocabulário*, e utiliza tais vetores para realizar a recuperação de textos que tratam de assuntos semelhantes, por exemplo. Palavras mais discriminantes de cada texto são extraídas e, em alguns casos, é atribuído um peso para cada uma delas de modo que as mais discriminantes tenham um peso maior.

Textos com conteúdo semelhante podem apresentar palavras específicas iguais, como termos técnicos, ou sinônimas e com uma determinada frequência. Fazendo-se uso de um vetor de características contendo a frequência de tais palavras, é possível classificar textos de acordo com o seu conteúdo ou recuperar textos relacionados a um determinado assunto. Para gerar tal representação, realiza-se a extração e a seleção de um conjunto de palavras, denominado vocabulário, de um grupo de textos de referência.

O vetor de características final de cada texto é calculado utilizando-se o vocabulário. Como mencionado, o vocabulário é constituído por um conjunto de palavras selecionadas e extraídas de textos utilizados como referência. Para tornar o processo de busca eficiente, é necessário extrair as características mais relevantes dos textos, ou seja, devem-se utilizar palavras que melhor discriminem um texto em particular ou conjunto de textos de assuntos relacionados, e eliminar aquelas que sejam comuns nos demais [Sivic e Zisserman 2009].

Definido o vocabulário, os textos passam, então, a serem representados pela frequência com que cada palavra pertencente ao vocabulário é encontrada no texto, tal como um histograma. Durante o processo de recuperação, os textos candidatos ao resultado de busca tem seus histogramas extraídos e comparados com os histogramas da base de dados utilizando-se de uma função de distância. Os histogramas com as menores distâncias são referentes aos textos mais semelhantes ao buscado.

Assim como em textos, imagens também sofrem com o problema do grande número de informações existentes, graças à popularização de dispositivos eletrônicos capazes de capturar imagens. Com esse crescimento, o processo de busca torna-se mais custoso surgindo, então, a necessidade da utilização de meios que facilitem sua recuperação, tal como a categorização das imagens, as quais são divididas em classes de acordo com o objeto presente nelas, por exemplo.

Em categorização de imagens, Csurka [Csurka et al. 2004] introduziu uma técnica baseada em BoW, na qual é utilizado o conjunto de descritores mais discriminativos de um conjunto de objetos de referência, para posteriormente calcular uma representação para cada imagem. Em sua proposta, busca-se elaborar uma técnica para realizar a categorização de imagens em geral, mantendo a simplicidade e eficiência computacional da BoW.

Dicionário de palavras visuais emprega as mesmas etapas de Dicionário de palavras, em que ambas obtém descritores a partir de um conjunto de referência, selecionam os mais discriminantes e, então, geram o dicionário a partir do qual serão criadas as novas representações dos dados a serem categorizados. A principal diferença entre elas está em como os descritores são obtidos, ou seja, devem empregar técnicas próprias para os tipos de dados com que trabalham.

O diagrama da Figura 3.1 apresenta, desde a etapa de formação do vocabulário visual, até a etapa na qual as imagens passam a ser identificadas a partir de histogramas de frequência das palavras do vocabulário construído. Cada uma das etapas apresentadas é descrita nas seções a seguir.

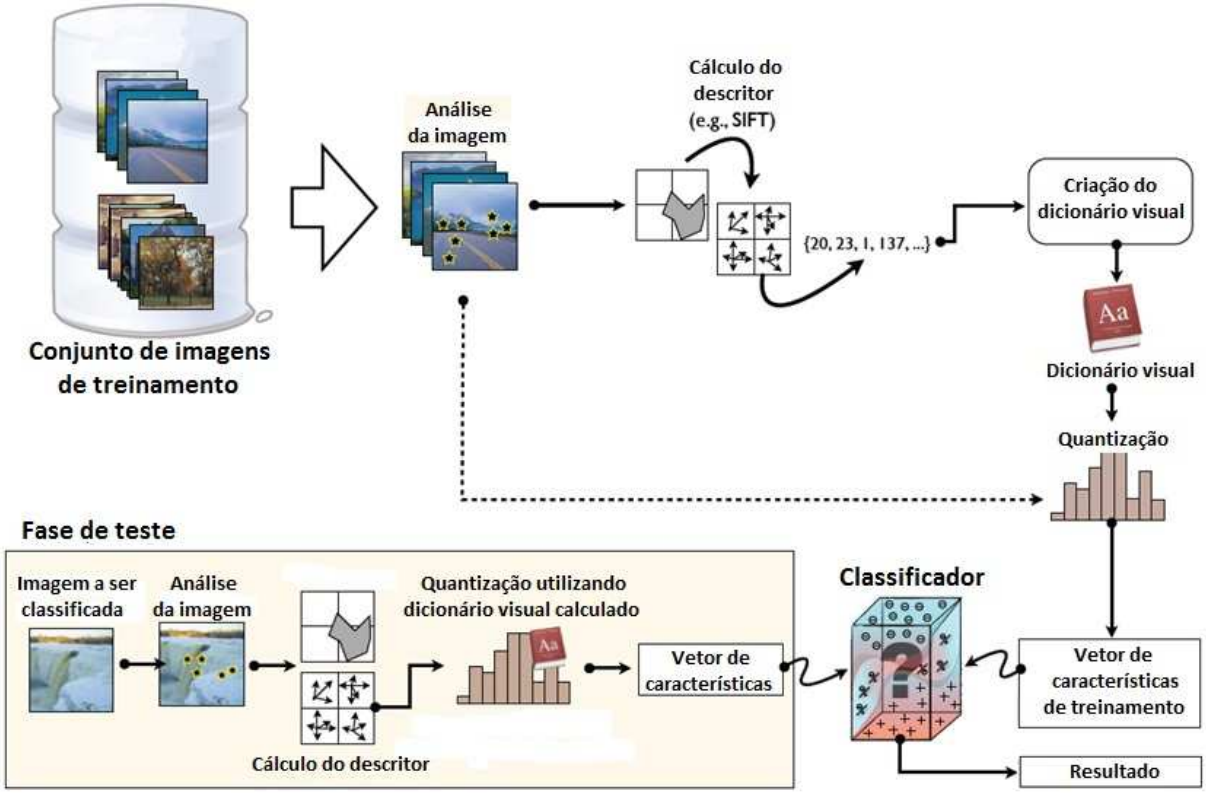


Figura 3.1: Etapas desenvolvidas em *BoVW*: realiza-se a análise da imagem utilizando filtros com o objetivo de identificar regiões estáveis; a partir das regiões identificadas são calculados os descritores e criado o dicionário visual; realiza-se a quantização gerando o vetor de características, o qual é utilizado para treinamento de uma rede neural. A fase de testes é semelhante à etapa de treinamento. Extraído e adaptado de [Papa e Rocha 2011].

3.2 Extração de Descritores

A extração de descritores (palavras visuais) é a primeira etapa e mais importante para obter-se o dicionário visual, já que a representação final das imagens será definida de acordo com o dicionário gerado e, quanto mais descriminativos os descritores, melhor será a taxa de acerto.

No mundo real, objetos apresentam-se em diferentes orientações e escalas, o que dificulta a categorização das imagens e pode prejudicar o desempenho do sistema de categorização caso tais transformações sejam desprezadas. Para contornar esse tipo de problema, são obtidos descritores invariantes à escala e rotação por meio de técnicas como *Scale Invariant Feature Transform* (SIFT) e *Speeded-Up Robust Feature* (SURF). A obtenção dos descritores é realizada por meio da análise das respostas de todas as regiões da imagem, em que as regiões mais estáveis são selecionadas. A análise envolve a utilização de filtros baseados na função gaussiana, o que seria equivalente a análise da imagem em diferentes escalas.

Após identificar as regiões estáveis, é realizado o cálculo dos descritores para cada uma delas considerando variações na escala e na orientação. A dimensão do descritor depende da técnica de extração aplicada, assim como o número de descritores obtidos depende da base de imagens.

3.3 Formação do Vocabulário Visual

O vocabulário é um item importante para BoVW, uma vez que os vetores de características das imagens são calculados a partir dele. Para realizar a categorização das imagens, assim como para a classificação de itens em geral, é necessário que a representação final proporcione um alto nível de distinção entre classes e que haja um alto nível de semelhança intra-classe para se obter um resultado final satisfatório.

O número de palavras visuais extraídas a partir de um dado conjunto de imagens de referência pode ser muito alto, devido as características da imagem, como alto número de regiões consideradas estáveis. O alto número de palavras visuais empregadas no vocabu-

lário visual pode prejudicar o desempenho do sistema, devido ao tempo necessário para realizar a comparação entre o vetor de características de uma dada imagem de entrada e daqueles das imagens armazenadas na base de dados.

Além disso, dentre as palavras visuais extraídas, estão aquelas que representam apenas pequenas variações de outras, ou seja, apresentam ruídos. A inclusão de tais palavras visuais também pode prejudicar o desempenho do sistema, pois essas podem não ser as mais discriminativas e, portanto, não acrescentam informação relevante ao vetor de características final.

Devido a esses motivos, é necessário realizar a seleção das palavras visuais que formarão o dicionário visual. Para essa tarefa podem ser aplicadas abordagens diversas, sendo *seleção aleatória* a mais simples delas. Nessa abordagem, cada palavra visual é identificada por um índice único e utiliza-se uma função que irá gerar uma quantidade n de números aleatórios, em que n representa a dimensão do vocabulário visual. Essa abordagem é a mais simples, mas pode não proporcionar resultados satisfatórios, uma vez que não é realizada uma análise para definir se a palavra visual é ou não discriminativa. Além disso, palavras visuais que representam um determinado padrão podem não ser selecionadas para compor o vocabulário visual.

Uma alternativa para a seleção de palavras visuais é aplicar algoritmos de agrupamento. Ao serem mapeadas em $\Re^n$, as palavras visuais formarão agrupamentos devido à semelhança entre elas. Aplicando-se um algoritmo de agrupamento, é possível determinar uma amostra mais representativa de cada agrupamento e, com isso, representar todos os padrões encontrados no conjunto de imagens de referência.

Apesar de serem métodos que podem proporcionar melhores resultados, muitos algoritmos de agrupamento necessitam da dimensão do vocabulário visual antes de realizarem os cálculos para definir as palavras visuais mais representativas. A necessidade de fornecer a dimensão como parâmetro inicial aparece como sendo a desvantagem desses algoritmos quando se deseja que o sistema de categorização atinja um desempenho satisfatório, pois muitas vezes não se sabe o tamanho ideal do dicionário.

A dimensão é importante, pois ela pode afetar o desempenho do sistema de maneira

negativa caso não seja bem definida já que ela define a representação final das imagens. A questão é que devido ao número de descritores, é possível criar vários dicionários visuais de dimensões distintas e com várias combinações de descritores, tornando a avaliação - realizar a quantização de cada imagem da base de dados e realizar as etapas de treinamento e teste do classificador - muito custosa. Por isso, utilizar algoritmos que criem o dicionário visual de maneira automática apresenta-se como uma vantagem.

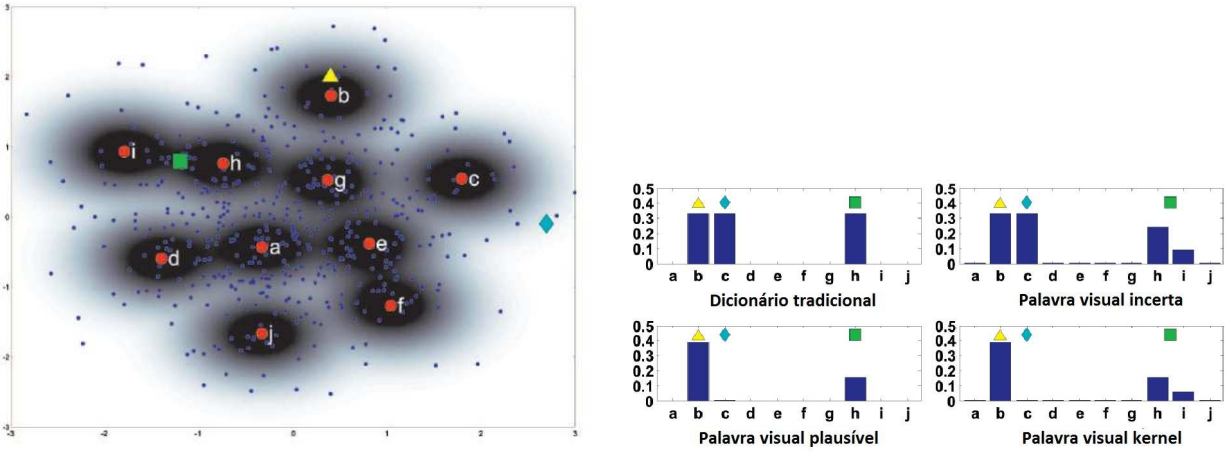
3.4 Histograma e Recuperação das Imagens

Cada imagem utilizada na etapa de formação do vocabulário, assim como aquelas que serão utilizadas durante o processo de recuperação, passarão a ser representadas na forma de um histograma. Dada uma imagem I e um vocabulário $V=\{w_1,w_2,...,w_n\}$, em que w_i denota uma palavra do vocabulário e n corresponde ao tamanho do dicionário, a nova representação de I será dada pela frequência de cada palavra visual w_i em I.

Descritores são extraídos de I e, então, realiza-se o cálculo da distância entre cada descritor e todas as palavras visuais do dicionário V por meio de uma métrica de distância, pode essa ser específica para o tipo de descritor extraído. A palavra w_i mais próxima terá o valor armazenado em sua respectiva posição no vetor de características, incrementado em 1 (uma) unidade.

No processo de recuperação ou classificação de imagens, os vetores de características podem ser calculados por meio da atribuição rígida [Philbin et al. 2008], em que o descritor é associado a uma única palavra visual ou podem-se atribuir pesos para cada um de seus componentes (atribuição leve), em que aqueles que são menos frequentes em diferentes imagens possuem peso maior ou pelo nível de semelhança entre o descritor e as palavras visuais, em que um descritor pode ser associado a mais de uma palavra visual [Boureau et al. 2010, Gemert et al. 2010].

Gemert *et al.* [Gemert et al. 2010] utiliza a chamada atribuição leve ao dicionário visual que tem apresentado resultados superiores à atribuição rígida. Além disso, seu trabalho introduz o emprego de ambiguidade que também apresenta uma melhora no desempenho de classificação. Por meio da ambiguidade, tem-se um maior nível de repre-



(a) (b)

Figura 3.2: Criação do vetor de características utilizando: (a) descritores de imagens de treinamento (pontos azuis), palavras visuais (pontos vermelhos) e descritores da imagem a ser quantizada (quadrado, triângulo e losango) e (b) vetores de características utilizando diferentes abordagens. Extraídos e adaptados de [Gemert et al. 2010].

sentação de cada imagem, enriquecendo o vetor de características e acarretando em um melhor desempenho na etapa de classificação [Gemert et al. 2010]. Ambiguidade pode ser modelada por diferentes métricas de semelhança, em que considera-se que um descritor pode ser semelhante a mais de uma palavra visual e tal semelhança pode ser representada por um intervalo entre 0 e 1. A Figura 3.2 apresenta um exemplo.

Neste trabalho, utilizamos os vetores de características sem a atribuição de qualquer peso para seus componentes (atribuição rígida) e a distância Euclidiana para definir a qual categoria pertence uma dada imagem de entrada.

3.5 Considerações Finais

O capítulo apresentou o método BoVW, o qual foi baseado no método BoW, descrevendo cada uma de suas etapas. O método BoVW cria um repositório de padrões extraídos de diferentes imagens que constituem a base de dados. As imagens, tanto as pertencentes à base de dados, quanto as imagens que serão identificadas, passam a ser representadas pela frequência com que cada palavra visual é encontrada nela. A partir dessa informação

de frequência, torna-se possível aplicar classificadores no processo de identificação, uma vez que todas as imagens são representadas por um vetor de características de mesma dimensão. O próximo capítulo apresenta os algoritmos de agrupamento utilizados neste trabalho.

4 Algoritmos de agrupamento

Este capítulo apresenta a fundamentação teórica dos algoritmos de agrupamento *OPF não-supervisionado* e *K-médias* utilizados durante os experimentos comparativos.

4.1 OPF não-supervisionado

OPF não-supervisionado é a versão não-supervisionada do classificador OPF que, assim como a versão supervisionada, também possui sua etapa de treinamento. A diferença entre elas é que OPF não-supervisionado trabalha com um conjunto de amostras não rotuladas, ou seja, a informação a qual classe cada uma das amostras pertence é ausente. Para realizar a etapa de treinamento encontra o número de classes existentes no conjunto baseando-se no parâmetro k_{max}, a ser explicado posteriormente. As classes são definidas por meio do cálculo de agrupamentos existentes utilizando um algoritmo de agrupamento baseado em grafos, em que cada amostra é representada como um nó. A seção a seguir descreve o algoritmo do OPF não-supervisionado com mais detalhes.

Seja Z uma base de dados tal que, para toda amostra $s \in Z$, existe um vetor de atributos $\vec{v}(s)$. Seja $d(s,t)$ a distância entre as amostras s e t no espaço de atributos. O problema fundamental na área de agrupamento de dados é identificar grupos de amostras em Z, sendo que amostras de um mesmo grupo deveriam representar algum nível de semelhança de acordo com algum significado semântico.

Primeiramente, definimos uma relação de adjacência A. Duas amostras s e t são ditas adjacentes quando satisfazem tal relação A, ou seja, $t \in A(s)$ ou $(s,t) \in A$. Por exemplo,

$$t \in A_1(s) \text{ se } d(s,t) \leq d_f \text{ ou} \tag{4.1}$$

$$t \in A_2(s) \text{ se } t \text{ é } k\text{-vizinho mais próximo de } s \text{ no espaço de atributos,} \tag{4.2}$$

em que d_f é o comprimento do maior arco em (Z, A_k), sendo este o par que define um grafo do tipo k-nn a partir de uma relação de adjacência do tipo A_2 e, posteriormente, do tipo A_3, Eq. 4.4.

Para o dado conjunto de dados Z, a técnica OPF não-supervisionado realizará o mapeamento de cada amostra $s \in Z$ como sendo um nó de um grafo. As arestas são ponderadas por $d(s,t)$ e os nós $s \in Z$ são ponderados por um valor de densidade probabilidade (fdp)

$$\rho(s) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2}|\mathcal{A}(s)|} \sum_{\forall t \in \mathcal{A}(s)} \exp\left(\frac{-d^2(s,t)}{2\sigma^2}\right), \tag{4.3}$$

em que A representa a relação de adjacência empregada e $\sigma = \frac{d_f}{3}$. A escolha deste parâmetro considera todos os nós para o cálculo da densidade, assumindo que uma função gaussiana cobre a grande maioria das amostras com $d(s,t) \in [0, 3\sigma]$. As amostras com valores máximos de fdp são definidas como sendo os nós centrais dos agrupamentos, porém, faz-se necessário a realização de algumas modificações tanto na relação de adjacência, quanto na função de propogação com o intuito de evitar problemas de super-segmentação do conjunto, apresentadas a seguir.

Relações de adjacência simétricas (Equação 4.1 por exemplo) resultam em relações de conectividade simétricas, entretanto A_2 na Equação 4.2 é uma relação de adjacência assimétrica. Dado que um máximo da fdp pode ser um subconjunto de amostras adjacentes com um mesmo valor de densidade, existe a necessidade da garantia da conectividade entre qualquer par de amostras naquele máximo. Assim, qualquer amostra deste conjunto de máximos pode ser representativa e alcançar outras amostras desse máximo e suas respectivas zonas de influência por um caminho ótimo.

Caso tal conectividade não exista, pode ocorrer super-segmentação uma vez que os

máximos vizinhos não irão pertencer ao mesmo agrupamento e, consequentemente, vários agrupamentos distintos serão formados. Por isso, é necessário uma modificação na relação de adjacência A_2, para que a mesma seja simétrica nos platôs de ρ com o intuito de calcular os clusters:

$$\begin{aligned} \text{se } t &\in \mathcal{A}_2(s), \\ s &\notin \mathcal{A}_2(t) \text{ e} \\ \rho(s) &= \rho(t), \text{ então} \\ \mathcal{A}_3(t) &\leftarrow \mathcal{A}_2(t) \cup \{s\}. \end{aligned} \tag{4.4}$$

Se tivéssemos uma amostra por máximo, formando um conjunto S (pontos grandes na Figura 4.1a), então a maximização da função f_1 resolveria o problema, ou seja:

$$\begin{aligned} f_1(\langle t \rangle) &= \begin{cases} \rho(t) & \text{se } t \in S \\ -\infty & \text{caso contrário} \end{cases} \\ f_1(\pi_s \cdot \langle s,t \rangle) &= \min\{f_1(\pi_s), \rho(t)\}. \end{aligned} \tag{4.5}$$

A função f_1 possui um termo de inicialização e um termo de propagação, o qual associa a cada caminho π_t o menor valor de densidade ao longo do mesmo. Toda amostra $t \in S$ define um caminho trivial $\langle t \rangle$ devido ao fato de não ser possível alcançar t a partir de outro máximo da fdp sem passar pelas amostras com valores de densidade menores que $\rho(t)$ (Figura 4.1a). As amostras restantes iniciam com caminhos triviais de valor $-\infty$ (Figura 4.1b), assim qualquer caminho oriundo de S possuirá valor maior. Considerando todos os caminhos possíveis de S a toda amostra $s \notin S$, o caminho ótimo $P^*(s)$ será aquele cujo menor valor de densidade seja máximo.

Visto que não temos os máximos da fdp, a função de conectividade precisa ser escolhida de tal forma que seus valores iniciais h definam os máximos relevantes da fdp. Para $f_1(\langle t \rangle) = h(t) < \rho(t)$, $\forall t \in Z$, alguns máximos da fdp serão preservados e outros serão alcançados por caminhos oriundos de outros máximos, cujos valores serão maiores do que

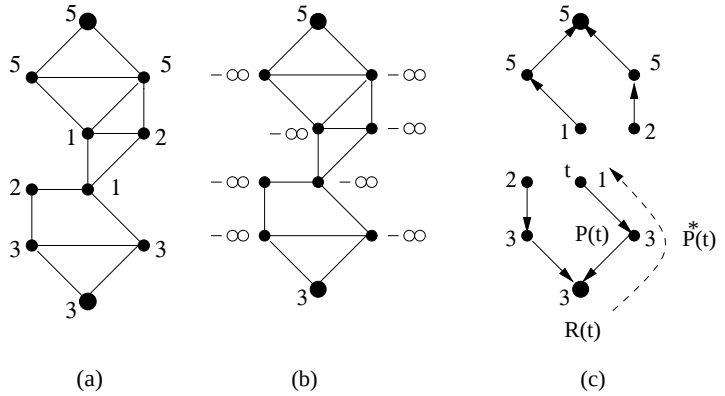


Figura 4.1: (a) Grafo cujos pesos dos nós são seus valores de fdp $\rho(t)$. Existem dois máximos com valores 3 e 5, respectivamente. Os pontos grandes indicam o conjunto de raízes S. (b) Valores de caminho triviais $f_1(\langle t \rangle)$ para cada amostra t. (c) Floresta de caminhos ótimos P para f_1 e os valores de caminho finais $V(t)$. O caminho ótimo $P^*(t)$ (linha tracejada) pode ser obtido percorrendo os predecessores $P(t)$ até a raiz $R(t)$ para cada amostra t. Extraído de [Rocha, Cappabianco e Falcão 2009]

seus valores iniciais. Por exemplo, se

$$\begin{aligned} h(t) &= \rho(t) - \delta, \\ \delta &= \min_{(s,t)\in\mathcal{A}|\rho(t)\neq\rho(s)} |\rho(t) - \rho(s)|, \end{aligned} \tag{4.6}$$

então todos os máximos de ρ serão preservados. Para altos valores de δ os domos da fdp com altura menor que δ não definirão zonas de influência.

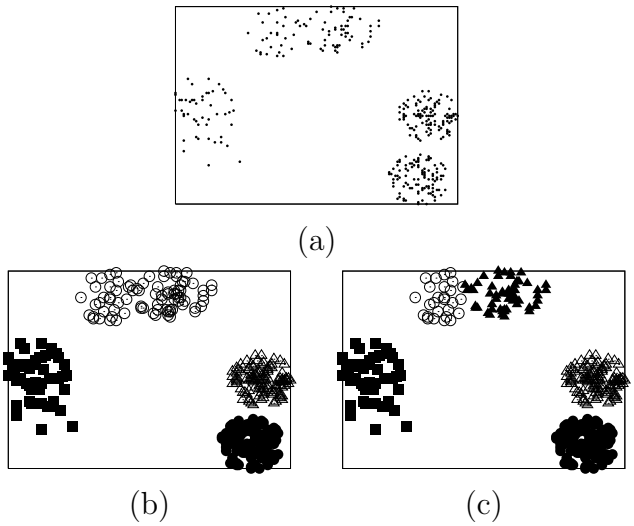
Outra questão a ser considerada é o parâmetro k, o qual define a relação A_k. O valor de k pode ser definido pelo usuário, assim como pode-se estabelecer um valor máximo para essa variável e, então, realizar a otimização do valor de k, como proposto por Rocha et al. [Rocha, Cappabianco e Falcão 2009]. Sua solução considera o corte mínimo no grafo provida pelos resultados do processo de não-supervisionado para $k^* \in [1, k_{max}]$, de acordo com a medida $C(k)$ sugerida por Shi e Malik [Shi e Malik 2000]:

$$C(k) = \sum_{i=1}^{c} \frac{W_i'}{W_i + W_i'}, \tag{4.7}$$

$$W_i = \sum_{\forall (s,t) \in \mathcal{A} | L(s)=L(t)=i} \frac{1}{d(s,t)}, \tag{4.8}$$

$$W_i' = \sum_{\forall (s,t) \in \mathcal{A} | L(s)=i, L(t)\neq i} \frac{1}{d(s,t)}, \tag{4.9}$$

em que $L(t)$ é o rotulo da amostra t, W_i' utiliza todos os pesos dos arcos entre o cluster i e os demais, e W_i utiliza todos os pesos dos arcos que pertencem ao cluster $i = 1, 2, \ldots, c$. A Figura 4.2a mostra um exemplo com $|Z| = 340$ amostras, as quais formam poucos clusters com diferentes concentrações de amostras no espaço de atributos 2D. Dependendo do valor de k escolhido, é possível encontrar até cinco agrupamentos de dados. Se $k_{max} \geq 150$, então o corte mínimo irá ocorrer quando todas as amostras estiverem agrupadas em um único cluster. O corte mínimo para $k_{max} = 100$ identifica quatro clusters com o melhor $k^* = 37$ (Figura 4.2b), e limitando a busca para $k_{max} = 30$, o corte mínimo identifica cinco clusters com melhor $k^* = 29$ (Figura 4.2c).



(a) (b) (c)

Figura 4.2: (a) Espaço de atributos com diferentes concentrações de amostras para cada cluster. É possível identificar diferentes quantidades de clusters dependendo do valor de k escolhido. Soluções interessantes são (b) quatro e (c) cinco clusters. Extraído de [Rocha, Cappabianco e Falcão 2009]

Segue abaixo o algoritmo do agrupamento de dados baseado em floresta de caminhos ótimos.

Algoritmo 1 – AGRUPAMENTO DE DADOS POR FLORESTA DE CAMINHOS ÓTIMOS

ENTRADA: Grafo (Z, A_{k^*}) e função ρ.
SAÍDA: Mapa de rótulos L, mapa de valores de caminho V, mapa de predecessores P.
AUXILIARES: Fila de prioridade Q, variáveis tmp e $l \leftarrow 1$.

1. **Para todo** $s \in Z$, **Faça** $P(s) \leftarrow nil$, $V(s) \leftarrow \rho(s) - \delta$, *insira* s *em* Q.
2. **Enquanto** Q *e não vazia,* **Faça**
3. *Remova de* Q *uma amostra* s *tal que* $V(s)$ *e maximo.*
4. **Se** $P(s) = nil$, **Então**
5. $L(s) \leftarrow l$, $l \leftarrow l + 1$, *e* $V(s) \leftarrow \rho(s)$.
6. **Para cada** $t \in A_{k^*(s)}$ *e* $V(t) < V(s)$, **Faça**
7. $tmp \leftarrow \min\{V(s), \rho(t)\}$.
8. **Se** $tmp > V(t)$, **Então**
9. $L(t) \leftarrow L(s)$, $P(t) \leftarrow s$, $V(t) \leftarrow tmp$.
10. *Atualize posição de* t *em* Q.

O Algoritmo 1 identifica uma raiz em cada máximo da fdp ($P(s) = nil$ na Linha 4 implica que $s \in S$), associa um rótulo distinto a cada raiz na Linha 5, e calcula a zona de influência (cluster) de cada raiz como sendo uma árvore de caminhos ótimos em P, tal que os nós de cada árvore recebem o mesmo rótulo que a sua raiz no mapa L (Linha 9). O algoritmo também retorna o mapa de valores de caminhos ótimos V e o mapa de predecessores P, sendo também mais robusto que o tradicional algoritmo de mean-shift [Cheng 1995], pois não depende de gradientes da fdp, utiliza um grafo k-nn e associa um rótulo para cada máximo, mesmo quando o máximo é composto por um componente conexo em (Z, A_{k^*}).

4.2 K-médias

K-médias foi proposto por MacQueen [MacQueen 1967] e é uma técnica simples para particionamento de um conjunto de dados n-dimensional Z em k aglomerados. Cada aglomerado é definido por um centróide c, o qual é o ponto central, assim como as amostras que a constituem são definidas a cada iteração. Durante as iterações, os centróides são ajustados, primeiramente, classificando cada amostra $a \in Z$ com o mesmo rótulo do centróide c_i mais próximo e, posteriormente, calculando suas novas posições por meio da média de cada aglomerado.

Os centróides, elementos centrais de cada agrupamento, são posicionados, inicialmente, de maneira aleatória no conjunto Z. Para cada amostra $a \in Z$, calcula-se a distância entre este e todos os centróides c_i, em que i define a classe do aglomerado, $i = 1, 2, \ldots, k$. Com isso, a amostra a pertencerá à classe do centróide c_i mais próximo.

Após todas as amostras a terem sido classificadas, realiza-se a atualização das posições dos centróides. Para cada aglomerado k_i, calcula-se o ponto médio e realiza-se a atualização da posição de c_i para o ponto médio de k_i. Os processos de classificação das amostras e de atualização dos centróides são realizadas até que a variação da posição dos centróides esteja abaixo de um determinado limiar. O Algoritmo 2 exemplifica o cálculo dos k centróides.

Algoritmo 2 – K-MÉDIAS

ENTRADA:	Um conjunto de amostras Z, número de agrupamentos k e limiar de distância l.
SAÍDA:	Conjunto de centróides S.
AUXILIARES:	Diferença d entre a posição anterior e atual dos centroides, centroide c_i, em que $i = 1, 2, \ldots, k$.

1. *Inserir k pontos de maneira aleatoria no espaço das amostras* $a \in Z$.
2. **Enquanto** $d > l$
3. | *Classificar as amostras a de acordo com a sua distância com relação aos centroides c.*
4. | *Substituir cada centroide* c_i *pelo ponto medio de seu respectivo agrupamento.*
5. | *Atualizar as posições de c em S.*

A Figura 4.3 apresenta a execução de cada uma das etapas.

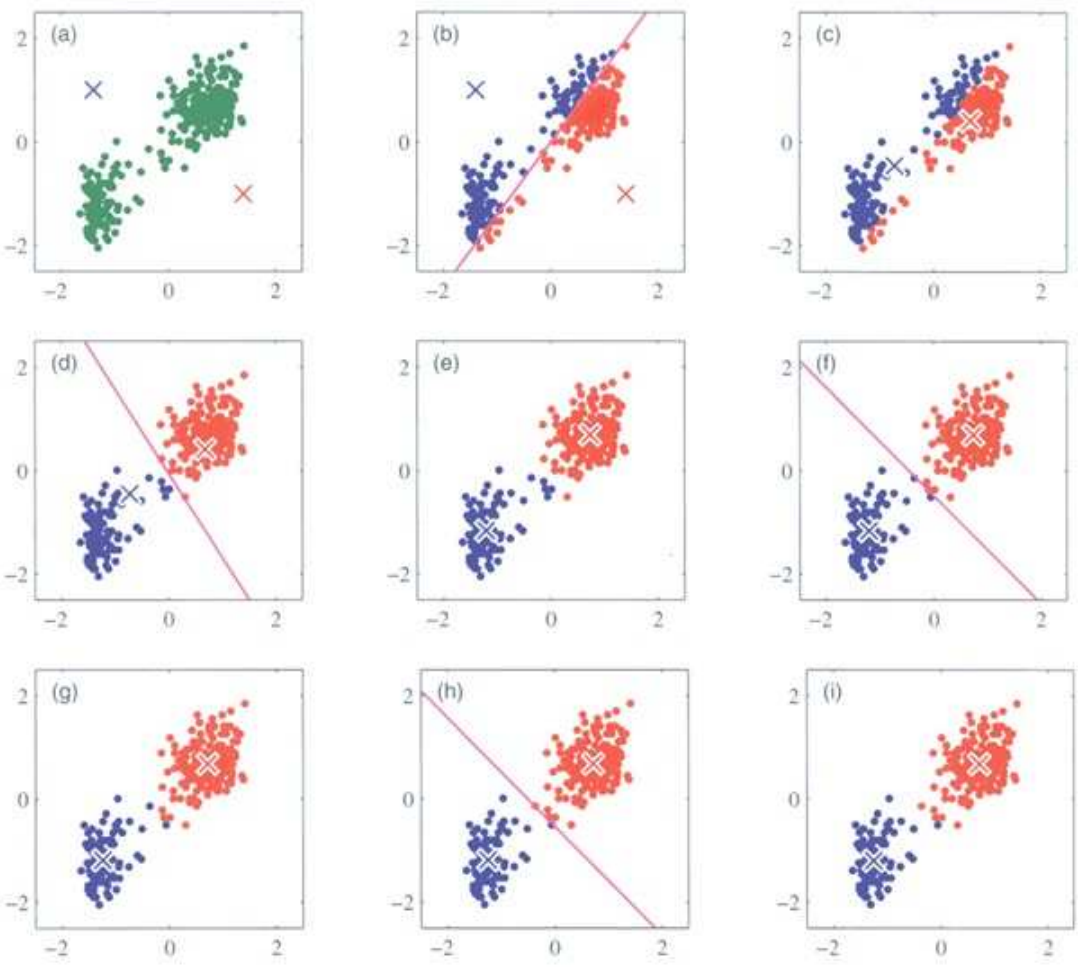


Figura 4.3: Etapas realizadas até encontrar os centróides finais. (a) posicionamento de dois centróides de maneira aleatória; (b), (d), (f) e (h) classificaç5ão de cada amostra, considerando a distância entre elas e os centróides; (c), (e), (g) e (i) atualização da posição dos centróides. Extraído de [Queensland].

4.3 Considerações Finais

Este capítulo apresentou os algoritmos de agrupamento aplicados nos experimentos. K-médias foi escolhido por ser um algoritmo simples e muito utilizado para realizar o

cálculo de agrupamentos de uma população. OPF não-supervisionado ainda não foi empregado para definir agrupamentos em populações. Seu conceito é semelhante ao utilizado pelo classificador OPF, em que os agrupamentos são definidos por meio de caminhos ótimos de um grafo e utilizando a fdp como peso de cada nó. Os classificadores utilizados na categorização das imagens são apresentados no próximo capítulo.

5 Classificador Bayesiano

Este capítulo apresenta o classificador Bayesiano, o qual foi empregado na etapa de avaliação dos métodos para geração dos dicionários visuais.

5.1 Teoria

É possível derivar uma estratégia de classificação ótima de padrões de tal modo que, na média, a probabilidade de se cometer erros seja mínima [Duda, Hart e Stork 2001].

A probabilidade de um determinado padrão x pertencer à classe ω_i é denotada por $p(\omega_i|x)$. Se o classificador de padrões decide que x pertence à classe ω_j quando na verdade ele pertence à classe ω_i, ocorre um erro ponderado por L_{ij}. Como o padrão x pode pertencer a M classes, o erro médio cometido ao se classificar x como sendo da classe ω_j é dado por:

$$r_j(x) = \sum_{k=1}^{M} L_{kj} p(\omega_k|x). \tag{5.1}$$

A equação 5.1 é denominada *erro médio condicional* na teoria da decisão e pode ser reescrita da seguinte forma:

$$r_j(x) = \frac{1}{p(x)} \sum_{k=1}^{M} L_{kj} p(x|\omega_k) P(\omega_k) \tag{5.2}$$

em que $p(x|\omega_k)$ é a função de densidade de probabilidade (fdp) dos padrões da classe ω_k. Como $\frac{1}{p(x)}$ é positivo e comum a todas as equações $r_j(x)$, para $j = 1, 2, \ldots, M$, a

equação 5.2 pode ser simplificada para:

$$r_j(x) = \sum_{k=1}^{M} L_{kj} p(x|\omega_k) P(\omega_k). \tag{5.3}$$

O classificador pode escolher, para cada padrão x com classificação desconhecida, M possíveis classes. Se ele calcula o erro médio $r_1(x), r_2(x), \ldots, r_M(x)$, para cada padrão x e o classifica como sendo pertencente à classe com menor erro, então o erro médio cometido durante a classificação de todos os padrões desconhecidos será mínimo.

O classificador que utiliza essa estratégia é conhecido como *Classificador Bayesiano.*

Em muitos problemas de reconhecimento, $L_{ij} = 1$ se $i \neq j$ e $Lij = 0$, se $i = j$. Nesses casos, a equação 5.3 pode ser reescrita da seguinte forma:

$$r_j(x) = p(x) - p(x|\omega_j) P(\omega_j) \tag{5.4}$$

Portanto, o classificador Bayesiano classifica o padrão x como pertencente à classe ω_i se:

$$p(x|\omega_i) P(\omega_i) > p(x|\omega_j) P(\omega_j), j = 1, 2, \ldots, M; i \neq j. \tag{5.5}$$

A função de decisão $d_j(x) = p(x|\omega_j) P(\omega_j)$ do classificador de Bayes depende do conhecimento prévio das FDPs $p(x|\omega_j)$ de cada classe ω_j, bem como das probabilidades de cada uma dessas classes ocorrer, $P(\omega_j)$.

A determinação das probabilidades de ocorrência das classes ω_j não constitui, em geral, um problema. Por exemplo, se todas as classes forem igualmente prováveis, então $P(\omega_j) = \frac{1}{M}$.

Um problema mais difícil consiste em determinar as FDPs de n variáveis das clas-

ses de padrões n-dimensionais. Na maioria das vezes, a única informação que se tem disponível sobre os padrões que se deseja classificar é um conjunto de amostras desses padrões. Nesses casos, podem ser adotadas dois tipos de técnicas para derivar um classificador baseado no conjunto de amostras: técnicas paramétricas e técnicas não paramétricas [Duda, Hart e Stork 2001].

Nas técnicas paramétricas supõe-se que os padrões a serem classificados apresentam FDPs conhecidas (por exemplo, supõe-se que elas sejam Gaussianas), e estimam-se os parâmetros que definem essas funções, utilizando-se os conjuntos de amostras.

As técnicas não-paramétricas não fazem suposições sobre as FDPs e tentam estimá-las diretamente a partir dos conjuntos de amostras dos padrões, ou tentam estimar as probabilidades à posteriori, $P(\omega_i|x)$, dos padrões para cada classe ω_i. Algumas técnicas de classificações não-paramétricas procuram transformar o espaço de atributos em outros espaços, nos quais podem-se utilizar técnicas paramétricas de classificação.

Na prática, é comum utilizar as técnicas paramétricas e supor que as fdp's $p(x|\omega_j)$ são funções Gaussianas. Assumida essa hipótese, basta estimar os parâmetros que definem a fdp de cada classe, utilizando para tanto conjuntos de amostras dessas classes. O resultado da classificação depende nesses casos de quão próxima da realidade está a hipótese assumida.

No caso n-dimensional, a densidade Gaussiana dos vetores pertencentes à j-ésima classe tem a forma:

$$p(x|\omega_j) = \frac{1}{(2\pi)^{n/2} \mid C_j \mid^{1/2}} exp\left[-\frac{1}{2}(x-\mu_j)^T C_j^{-1}(x-\mu_j)\right] \tag{5.6}$$

em que μ_j é o vetor médio e C_j é a matriz de covariância, definidos como:

$$\mu_j = E_j\{x\} \tag{5.7}$$

e

$$C_j = E_j\{(x - \mu_j)(x - \mu_j)^T\} \tag{5.8}$$

em que $E\{.\}$ denota o valor esperado, n é a dimensão do vetor, e $\mid C_j \mid$ é o determinante da matriz C_j.

Os parâmetros μ_j e C_j que definem a fdp Gaussiana da classe ω_j podem ser estimados tomando-se um conjunto de amostras de padrões x pertencentes a essa classe, utilizando-se a seguinte aproximação:

$$\mu_j = \frac{1}{N_j} \sum_{x \in \omega_j} x \tag{5.9}$$

e

$$C_j = \frac{1}{N_j} \sum_{x \in \omega_j} (xx^T - \mu_j \mu_j^T) \tag{5.10}$$

em que N_j é o número de amostras da classe ω_j.

A Figura 5.1 mostra um exemplo em que existem duas classes de padrões unidimensionais, ω_1 e ω_2, cujas funções de densidade de probabilidades são dadas por duas funções Gaussianas. O ponto x_0 é o limite de decisão utilizado pelo classificador, caso as duas classes sejam igualmente prováveis de ocorrer, ou seja, $P(\omega_1) = P(\omega_2)$.

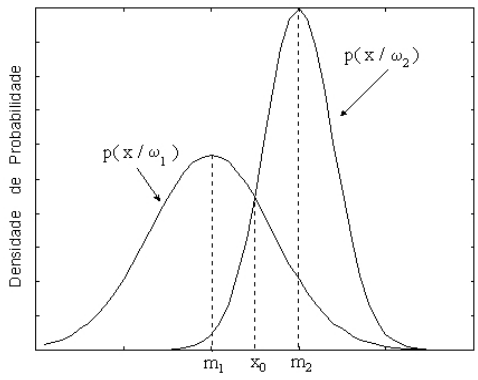


Figura 5.1: Funções unidimensionais de densidades de probabilidades para duas classes de padrões, ω_1 e ω_2. Extraído de [Duda, Hart e Stork 2001].

5.2 Considerações Finais

Este capítulo introduziu o classificador Bayesiano, o qual é um classificador estatístico que classifica uma dada amostra de entrada de acordo com a probabilidade desta pertencer a uma determinada classe. A metodologia e os resultados experimentais são apresentados no próximo capítulo.

6 Metodologia e Experimentos

No presente trabalho, foram utilizadas 3 bases de dados públicas contendo imagens de diversos objetos:

- ETH80[1]: contém 328 imagens divididas em 8 classes;
- Fish cells [Papa et al. 2011]: possui 4 classes com 20 imagens cada;
- Tropical Fruits[2]: 1.125 imagens divididas em 15 classes, com 75 imagens cada.

A metodologia empregada no desenvolvimento dos experimentos segue as etapas realizadas pela técnica BoVW. Primeiramente, são extraídos os descritores SIFT e SURF de cada imagem de cada base, obtendo-se assim 2 conjuntos de descritores para cada base. O número de descritores obtidos varia de acordo com os tipos de imagens utilizadas, em que aquelas que possuem um maior número de bordas e contornos proporcionarão um maior número de descritores, uma vez que essas regiões são estáveis com relação à escala. Vale lembrar que os 2 conjuntos de descritores obtidos são utilizados separadamente, ou seja, não é realizada qualquer tipo de operação de forma a fundir os descritores SIFT e SURF.

Após a extração dos descritores, realiza-se o cálculo dos dicionários visuais. Como mencionado, é a partir dos dicionários visuais que se obtém a base de dados final que será utilizada nas etapas de treinamento e teste dos classificadores, por meio do processo de quantização. Com o intuito de estudar a influência do tamanho do dicionário na taxa de acerto dos classificadores, foram computados 10 dicionários visuais de tamanhos distintos para cada conjunto de descritores e para cada base.

[1]http://www.vision.ee.ethz.ch/~bleibe/data/datasets.html <Acesso em: nov. 2011>

[2]http://www.ic.unicamp.br/~rocha/pub/downloads/tropical-fruits-DB-1024x768.tar.gz <Acesso em: nov. 2011>

Anterior ao cálculo dos dicionários, foi realizada a validação cruzada por 5-*folds* em cada conjunto de descritores. O propósito dessa idéia é criar diferentes conjuntos de treinamento por meio das possíveis combinações dos conjuntos calculados, onde 4 *folds* são utilizados para treinamento e o restante para classificação. Portanto, têm-se 5 conjuntos de treinamento distintos para cada conjunto de descritores. A partir de cada conjunto de treinamento, são calculados os 10 dicionários visuais utilizando 3 técnicas distintas (seleção aleatória, K-médias e OPF não-supervisionado). Por isso, cada técnica será responsável por gerar 50 dicionários para cada conjunto de descritores e para cada base, ou seja, 10 dicionários para cada combinação de *folds*.

A avaliação das técnicas nessa etapa é realizada por meio da medição do tempo necessário para gerar cada dicionário visual. Deseja-se que, além dos dicionários visuais sejam calculados no menor tempo possível, que esses proporcionem taxas de acerto satisfatórias com o menor número de palavras visuais possível.

Por que utilizar o 5-fold no conjunto de descritores antes de realizar o cálculo dos dicionários visuais, ao invés de utilizar todo o conjunto? Ao não realizar essa etapa, estaremos considerando que qualquer descritor pertencente ao conjunto pode ser selecionado para formar o dicionário visual. Com isso, alguns descritores que foram selecionados podem pertencer às imagens do conjunto de teste podendo, de alguma maneira, influenciar na taxa de acerto dos classificadores.

Uma vez gerados os dicionários visuais, realiza-se, então, o processo de quantização. O processo de quantização é necessário para gerar vetores de características de mesma dimensão para todas as imagens, possibilitando o uso de classificadores. A quantização é realizada tanto no conjunto de treinamento, quanto no conjunto de classificação, utilizando o mesmo dicionário visual, resultando em dois conjuntos quantizados. Portanto, cada dicionário irá gerar 5 conjuntos de treinamento e 5 conjuntos de teste (referentes ao número de *folds*). Uma vez obtidos os conjuntos de treinamento e teste, inicia-se, então os experimentos de classificação.

Os experimentos são constituídos por uma única execução de cada par de conjuntos (treinamento e teste). Para cada execução são computados os tempos de treinamento e classificação, assim como a taxa de acerto para o dado par. Ao final das execuções, é

calculado o valor médio de cada métrica. A taxa de acerto utilizada neste trabalho é dada por:

$$Acc = \frac{\sum_{i=1}^{n} ta_i}{n}, \tag{6.1}$$

em que ta_i representa a taxa de acerto para cada classe i e n representa o número total de classes.

Dentre os classificadores existentes, este trabalho utilizou o classificador Bayesiano Naive por ser um dos mais tradicionais na literatura e por sua fácil implementação. As Seções 6.1 e 6.2 apresentam, respectivamente, os resultados de tempo e taxa de acerto para cada uma das técnicas empregadas.

6.1 Resultados de Tempo de Cálculo dos Dicionários Visuais

Os gráficos das Figuras 6.1 à 6.6 apresentam os resultados médios e desvio padrão de tempo de cálculo dos dicionários por cada técnica utilizando descritores SURF e SIFT. No eixo das abscissas estão representados os valores de k_{max} utilizados pelo algoritmo OPF não-supervisionado. As 10 dimensões resultantes foram empregadas nos algoritmos K-médias e Seleção Aleatória para que pudesse ser realizada uma comparação justa entre o desempenho de cada técnica. Pelo fato de serem utilizados 5 conjuntos distintos para treinamento e testes, as dimensões para um mesmo valor de k_{max} também foram distintos. A Tabela 6.1 apresenta as dimensões mínima e máxima para cada valor de k_{max} empregado em cada base e cada descritor. Vale ressaltar que as dimensões abaixo apresentadas também foram empregadas nos algoritmos K-médias e Seleção aleatória.

A partir dos resultados, nota-se que o tempo necessário para calcular dicionários de menor dimensão é maior quando utiliza-se o algoritmo OPF não-supervisionado, diferente do que ocorre com K-médias e Seleção Aleatória. O fato do K-médias ter tempo proporcional à dimensão é que o número de distâncias entre as amostras e os centróides é menor. Já para o algoritmo OPF não-supervisionado, um número menor de agrupamentos indica um valor alto de k_{max}. Isso significa que cada protótipo realizará um número maior de cálculos para definir os caminhos ótimos na floresta.

	ETH80		Fish Cells		Tropical Fruits	
k_{max}	SIFT	SURF	SIFT	SURF	SIFT	SURF
1	9.950-10.253	5.067-5.279	609-656	417-466	147.608-151.154	29.157-29.902
11	239-266	76-93	32-38	15-20	127-3.872	419-428
21	88-106	41-54	12-21	5-11	1.371-1.478	230-239
31	41-53	26-32	8-14	2-4	721-776	154-179
41	29-46	19-31	4-11	2-3	446-495	113-140
51	22-25	16-25	4-7	2-3	316-372	100-110
61	16-24	13-21	3-5	1-2	240-281	76-100
71	11-21	13-21	2-4	1-2	184-243	75-84
81	7-20	11-19	1-4	1-2	149-185	58-77
91	10-14	12-15	1-4	1-2	145-172	45-72

Tabela 6.1: Dimensões máximas e mínimas dos dicionários para cada valor de k_{max}.

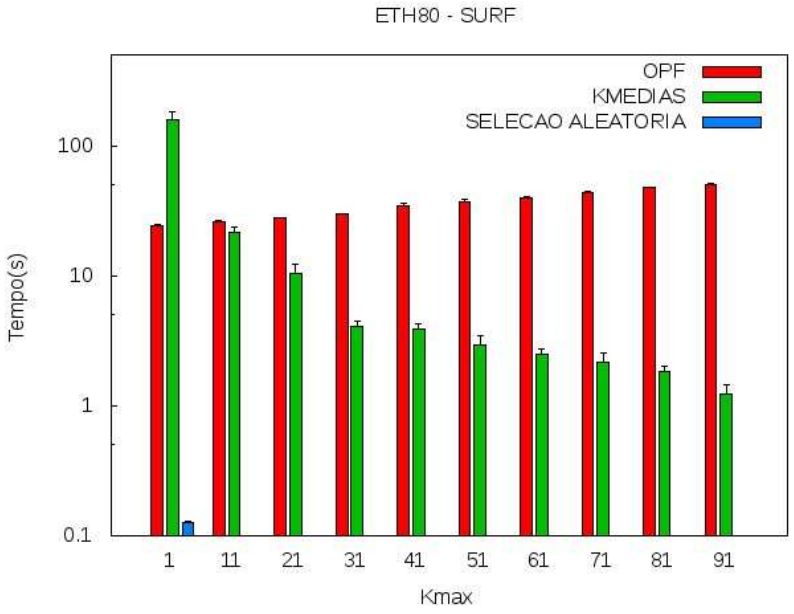


Figura 6.1: Tempos ETH80 utilizando descritores SURF.

Outro ponto a ser destacado com relação aos tempos ocorre na base Tropical Fruits. Apesar de apresentar uma oscilação de tempo, a técnica OPF não-supervisionado apresenta um tempo quase constante com relação às demais técnicas. Uma característica dessa base é o número de descritores extraídos ser até quase 235 vezes superior ao número de descritores com que as técnicas estão trabalhando nas outras 2 bases. O comportamento quase constante é algo que deve ser mais explorado para avaliar se tal comportamento se repete para outras bases com as mesmas características. A Tabela 6.2 apresenta o número de descritores calculados para cada uma das bases.

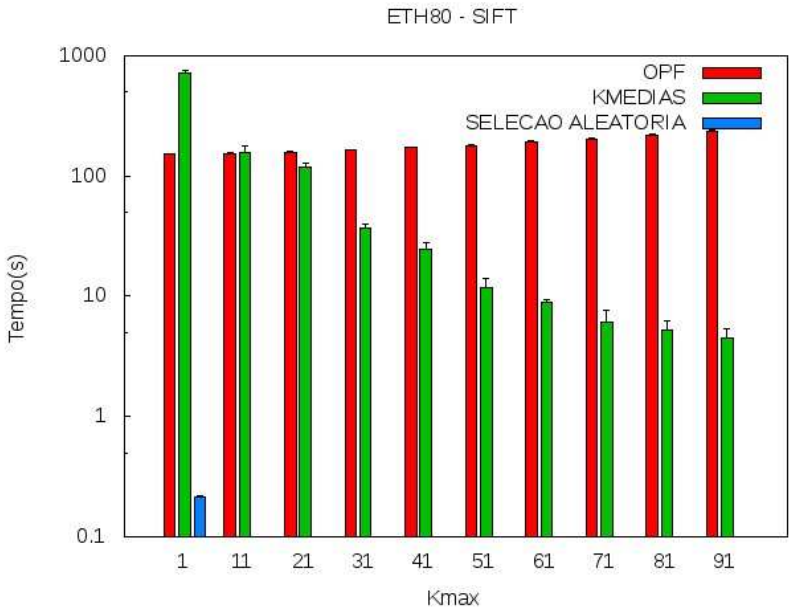


Figura 6.2: Tempos ETH80 utilizando descritores SIFT.

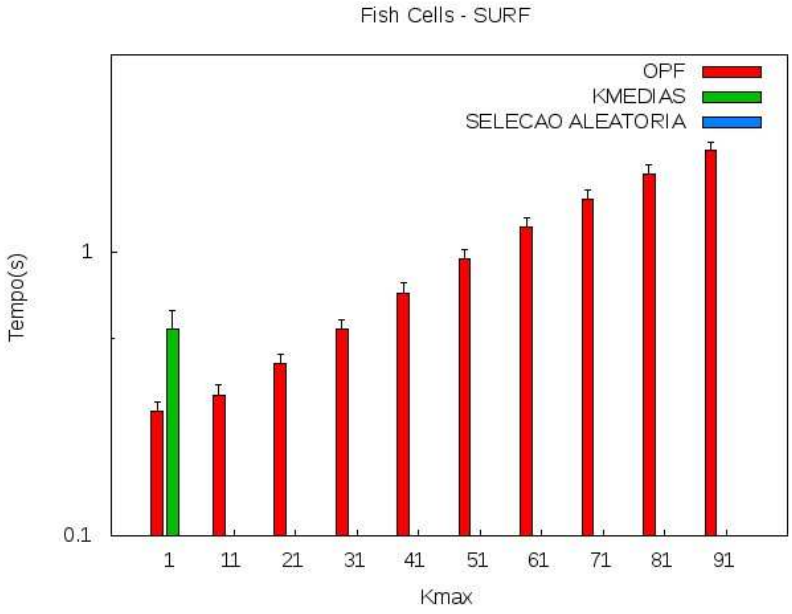


Figura 6.3: Tempos Fish Cells utilizando descritores SURF.

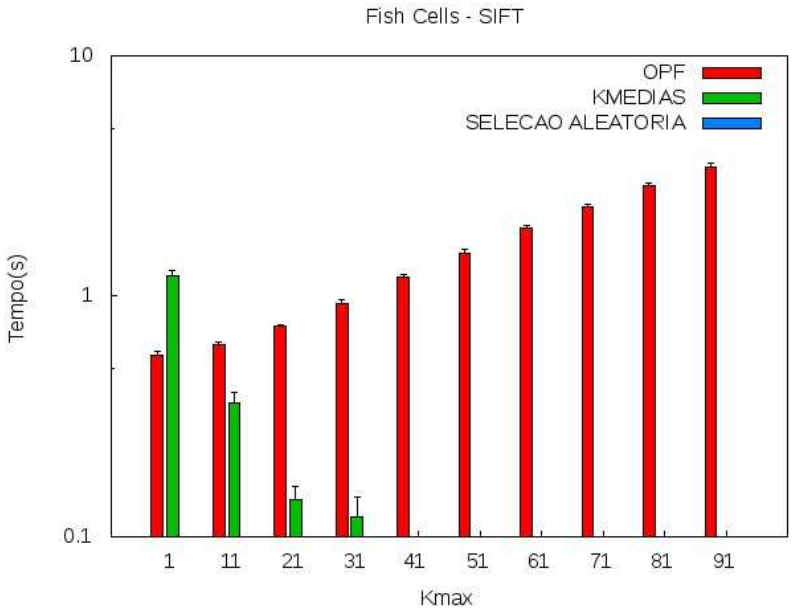


Figura 6.4: Tempos Fish Cells utilizando descritores SIFT.

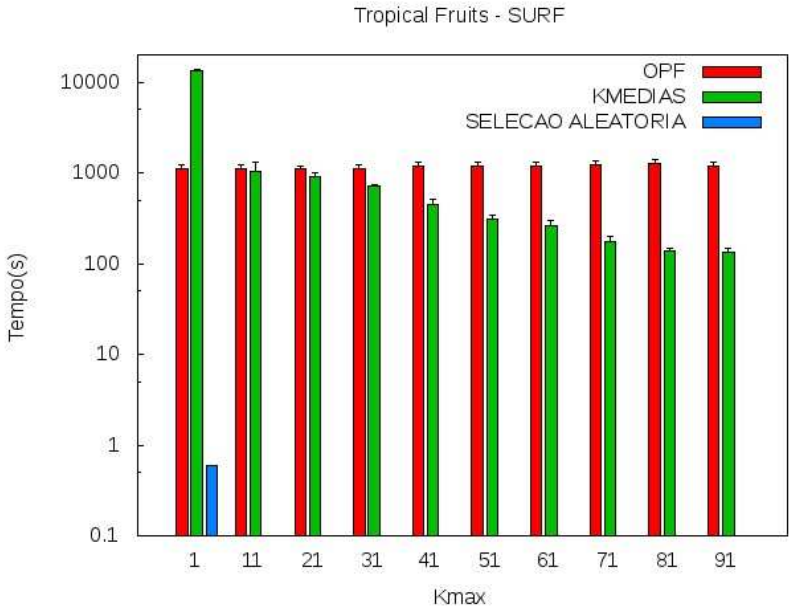


Figura 6.5: Tempos Tropical Fruits utilizando descritores SURF.

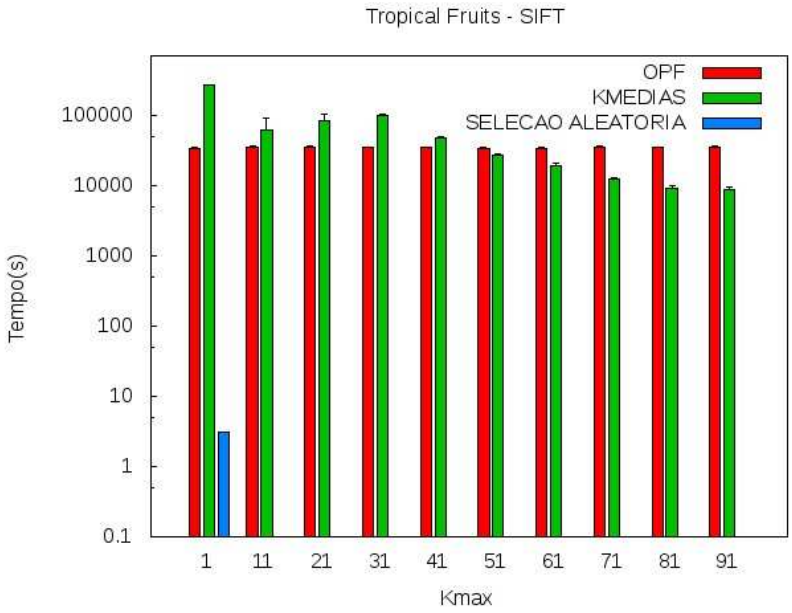


Figura 6.6: Tempos Tropical Fruits utilizando descritores SIFT.

Base	**SIFT**	**SURF**
ETH80	22.764	11.834
Fish Cells	1.429	1.005
Tropical Fruits	335.422	66.795

Tabela 6.2: Número de descritores obtidos em cada base.

6.2 Resultados de Taxa de Acerto

Os gráficos das Figuras 6.7 à 6.12 apresentam as taxas de acerto médias de para cada conjunto de dicionários para cada descritor em cada base, utilizando o classificador Bayesiano.

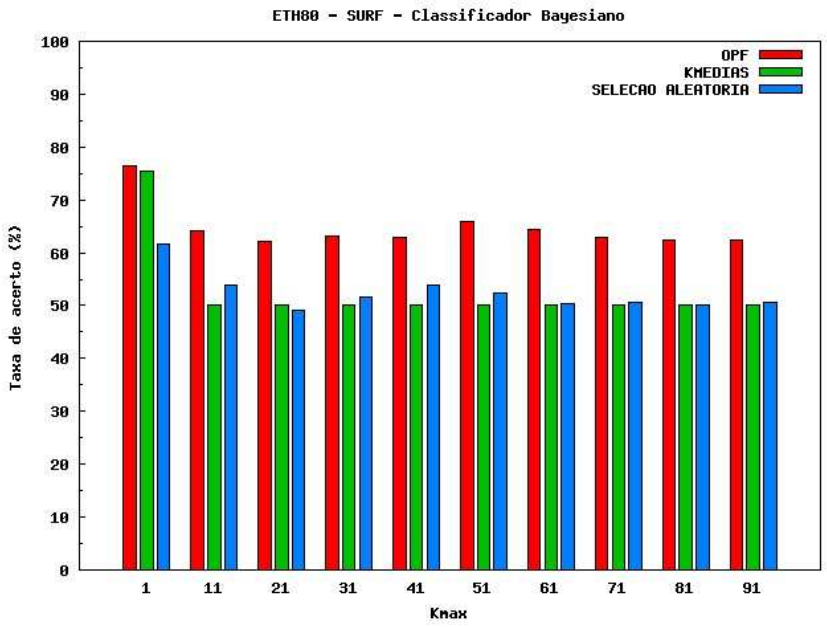


Figura 6.7: Taxa de acerto ETH80 utilizando descritores SURF e classificador Bayesiano.

Os resultados para a base ETH80 demonstram que a taxa de acerto proporcionado pelos dicionários gerados pelo OPF não-supervisionado utilizando descritor SURF foram superiores às taxas em todos os casos. Contudo, para descritores SIFT, seu desempenho foi abaixo das técnicas K-médias e Seleção Aleatória, em que o último apresentou melhores resultados médios em alguns casos, mas havendo empate em quase todas as situações. Diferentemente do que ocorreu na base ETH80 utilizando descritores SURF, na base Fish Cells, com o mesmo descritor, há um destaque para os desempenhos das técnicas OPF não-supervisionado e Seleção Aleatória. Já utilizando descritores SIFT, o desempenho das 3 técnicas foi semelhante ao ocorrido na base ETH80 e também ao que ocorreu na base Tropical Fruits. Novamente com descritores SURF, mas na base Tropical Fruits, destacam-se os desempenhos das técnicas OPF não-supervisionado e Seleção Aleatória com as melhores taxas de acerto e mínima diferença entre elas.

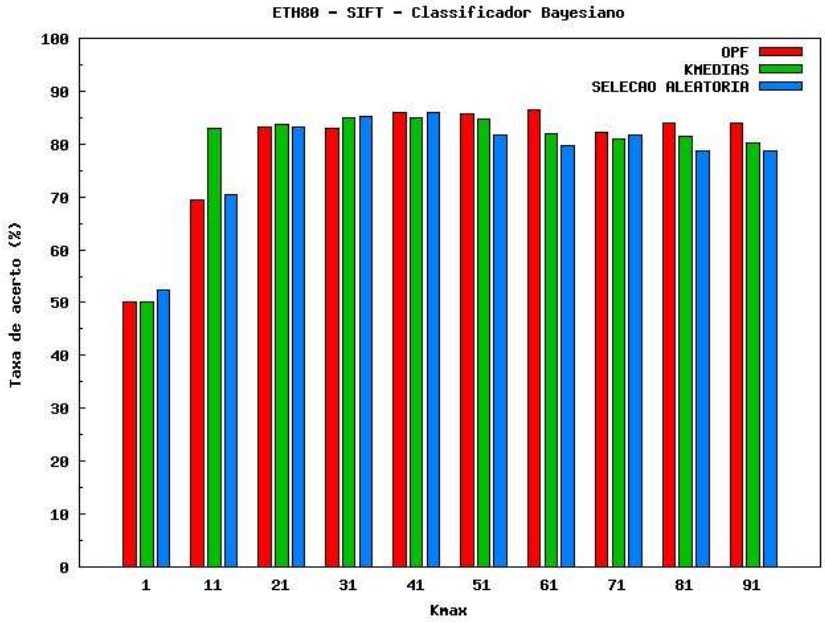


Figura 6.8: Taxa de acerto ETH80 utilizando descritores SIFT e classificador Bayesiano.

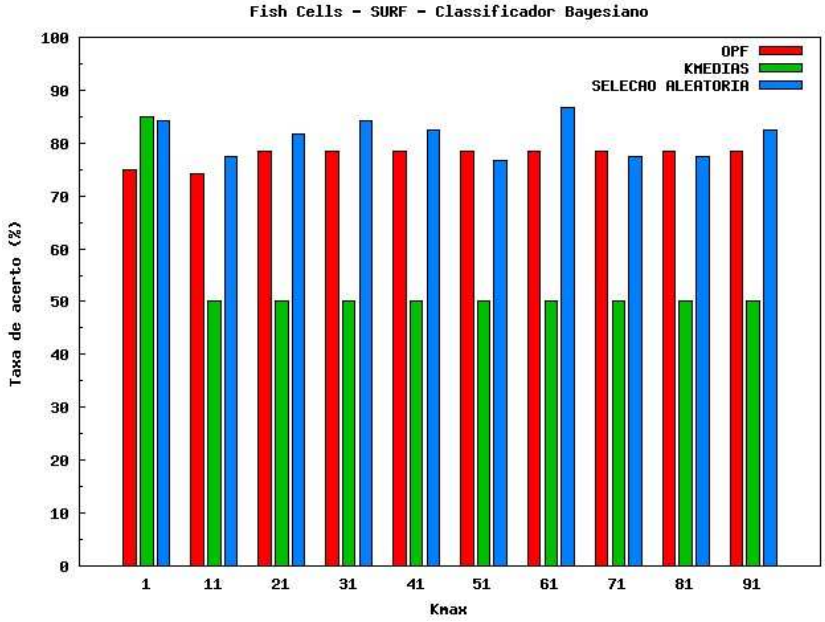


Figura 6.9: Taxa de acerto Fish Cells utilizando descritores SURF e classificador Bayesiano.

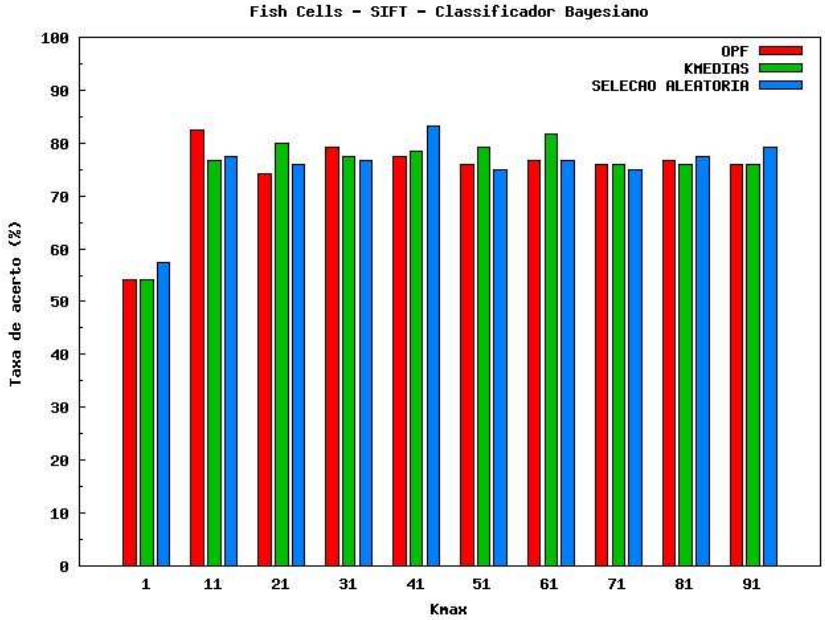


Figura 6.10: Taxa de acerto Fish Cells utilizando descritores SIFT e classificador Bayesiano.

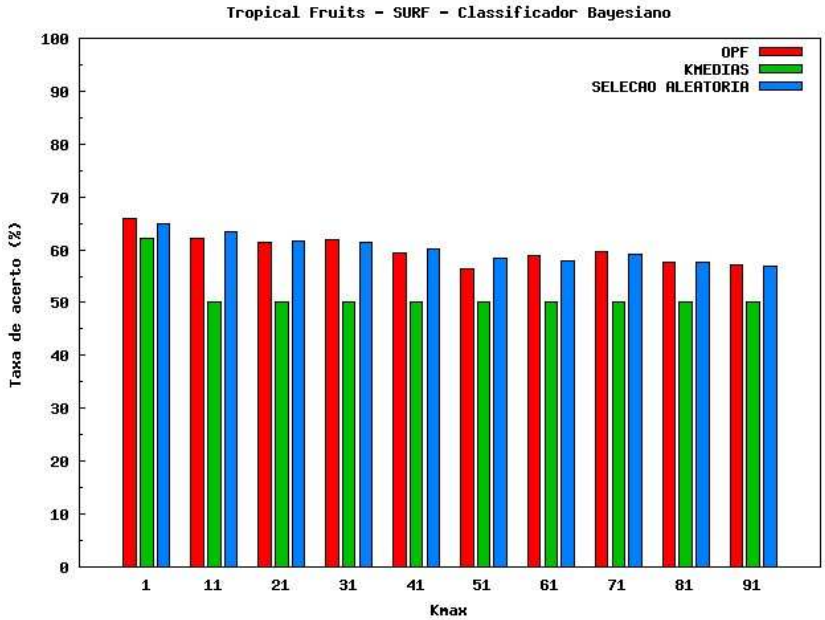


Figura 6.11: Taxa de acerto Tropical Fruits utilizando descritores SURF e classificador Bayesiano.

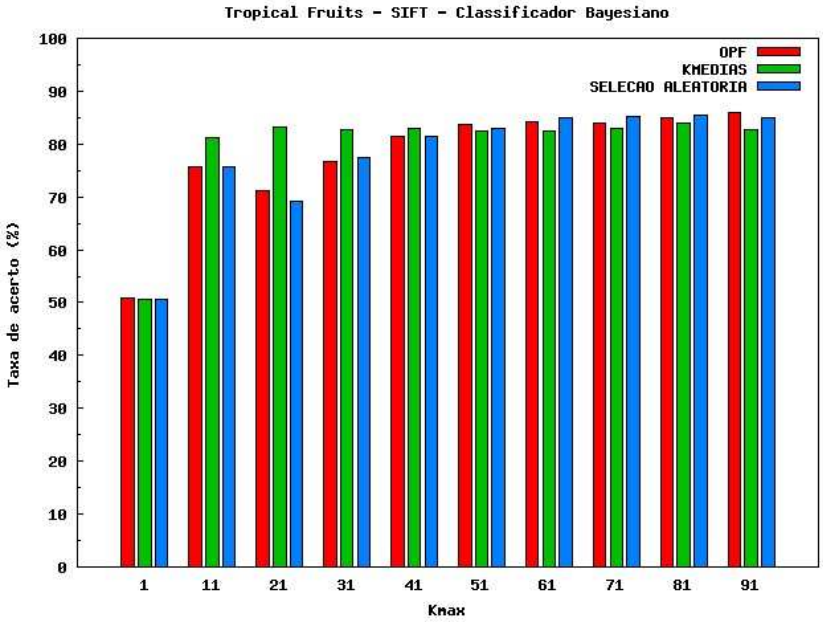


Figura 6.12: Taxa de acerto Tropical Fruits utilizando descritores SIFT e classificador Bayesiano.

7 Conclusões

O presente trabalho teve como objetivo avaliar a técnica de agrupamento OPF com relação à taxa de acerto proporcionada ao sistema e ao seu tempo de execução aplicada à seleção de descritores a serem utilizados no dicionário visual, sem a necessidade de definir a dimensão do conjunto a ser calculado. A avaliação é realizada por meio de experimentos comparando seu desempenho com as outras técnicas utilizadas neste projeto, como seleção aleatória e K-médias.

A seleção aleatória é implementada como sendo, basicamente, uma função que gera um conjunto de números aleatórios, em que cada número representa o índice de cada palavra visual extraída do conjunto de imagens de referência. Sua vantagem é a simplicidade do método e a rapidez com que os dicionários são calculados. Sua desvantagem é justamente a maneira como as palavras visuais são selecionadas, onde palavras visuais consideradas discriminativas podem não ser selecionadas.

O classificador K-médias, por outro lado, realiza uma melhor seleção das palavras visuais, uma vez que esta técnica define as palavras visuais como sendo as amostras centrais de cada agrupamento. O número de palavras visuais é definido antes da realização do cálculo do dicionário visual. Supondo um dicionário de dimensão k, o algoritmo irá definir, de maneira aleatória, k amostras como sendo as amostras centrais (centróides). A cada iteração, as demais amostras são classificadas de acordo com a sua distância à cada centróide e atualização da posição dos mesmos por meio do cálculo do ponto médio de cada agrupamento. O processo é realizado até que não haja mudança significativa na posição dos centróides.

O algoritmo OPF não-supervisionado realiza a seleção das palavras visuais sem a

necessidade da dimensão antes de realizar os cálculos. Para isso, calcula-se a *fdp* para cada nó pertencente ao grafo, assim como a distância entre eles. Os nós com os maiores valores de *fdp* são definidos como sendo os prótotipos (amostras centrais de cada agrupamento), os quais oferecem caminhos de custo ótimo para os demais nós formando, então, os agrupamentos.

As técnicas empregadas de seleção de palavras visuais foram avaliadas utilizando-se bases de imagens variadas, juntamente com descritores (SIFT e SURF) e o classificador Bayesiano. Avaliou-se o ganho proporcionado por cada técnica, assim como o seu desempenho com relação ao tempo necessário para realizar os cálculos dos dicionários visuais. Durante os experimentos foram utilizados dicionários de dimensões distintas para avaliar o impacto causado na taxa de acerto.

Os experimentos também avaliaram se K-médias e OPF não-supervisionado podem proporcionar melhores taxas de acerto com relação à seleção aleatória, uma vez que os primeiros realizam cálculos para encontrar grupos de padrões semelhantes e selecionam as amostras centrais de cada agrupamento que, hipoteticamente, seriam as amostras mais discriminantes. Por meio dos resultados experimentais, observou-se um ganho na taxa de acerto em poucos casos, em que na maioria das situações os desempenhos de acurácia são semelhantes.

Também por meio dos resultados obtidos, notou-se que dicionários com alta dimensão não necessariamente proporcionam melhores taxas de acerto. Isso ocorreu nos seguintes casos:

- ETH80 com descritor SIFT: 86.56% de taxa de acerto com dicionários variando entre 29 e 46 palavras visuais e $k_{max} = 41$;
- Fish Cells com SIFT: 83.33% com dicionários variando entre 8 e 14 palavras visuais e $k_{max} = 31$;
- Fish Cells com SURF: 86.67% de taxa de acerto com dicionário de dimensão variando entre 1 e 2 palavras visuais ($k_{max} = 61$), contra 85.00% com dicionários de dimensão variando entre 417 e 466 palavras visuais e $k_{max} = 1$;

- Tropical Fruits com SIFT: 85.00% com dicionários de dimensão variando entre 145 e 172 palavras visuais e $k_{max} = 91$, em que os maiores dicionários possuem entre 147.608 e 151.154 palavras visuais.

Como apresentado nos resultados, o algoritmo OPF não-supervisionado levou mais tempo para calcular dicionários de menor dimensão e menos tempo para calcular dicionários de maior dimensão; o oposto ocorre com K-médias. Também observou-se que em situações com alto número de descritores, o desempenho de tempo do OPF não-supervisionado mostrou-se aproximadamente constante.

A partir dos resultados, verificou-se que OPF não-supervisionado proporciona taxas de acerto próximas das outras técnicas avaliadas podendo ser uma alternativa para realizar a etapa de cálculo dos dicionários visuais, substituindo técnicas como K-médias e seleção aleatória, uma vez que OPF não-supervisionado é capaz de calcular os dicionários sem a necessidade do parâmetro dimensão do dicionário visual. Sua principal desvantagem, é que o tempo necessário para gerar o dicionário é inversamente proporcional ao tamanho do dicionário, ou seja, quanto maior o dicionário, maior o tempo.

8 Trabalhos Publicados

- Afonso, Luis C. S. ; Papa, João P. ; Papa, Luciene P. ; Marana, Aparecido N. ; Rocha, Anderson R. . Automatic Visual Dictionary Generation Through Optimum-Path Forest. In: IEEE International Conference on Image Processing, 2012, Orlando. 2012.

- Afonso, Luis C. S. and Papa, João P. and Marana, Aparecido Nilceu and Poursaberi, Ahmad and Yanushkevich, Svetlana N. A fast large scale iris database classification with Optimum-Path Forest technique: A case study. IJCNN. 1-5, IEEE. 2012.

- Afonso, Luis C. S. and Papa, João P. and Marana, Aparecido Nilceu and Poursaberi, Ahmad and Yanushkevich, Svetlana N. and Gavrilova, Marina L. Optimum-Path Forest Classifier for Large Scale Biometric Applications. EST. editor(s) Stoica, Adrian and Zarzhitsky, Dimitri and Howells, Gareth and Frowd, Charlie D. and McDonald-Maier, Klaus D. and Erdogan, Ahmet T. and Arslan, Tughrul. 58-61, IEEE Computer Society, 2012.

- Mansano, A.; Matsuoka, J.A.; Afonso, L.C.S.; Papa, J.P.; Faria, F.; da S Torres, R.; , "Improving Image Classification through Descriptor Combination,"Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on , vol., no., pp.324-329, 22-25 Aug. 2012

Referências

[Bay et al. 2008]BAY, H. et al. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, Elsevier Science Inc., New York, NY, USA, v. 110, n. 3, p. 346–359, jun. 2008. ISSN 1077-3142. Disponível em: <http://dx.doi.org/10.1016/j.cviu.2007.09.014>.

[Bosch, Zisserman e Muñoz 2008]BOSCH, A.; ZISSERMAN, A.; MUÑOZ, X. Scene classification using a hybrid generative/discriminative approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, Washington, DC, USA, v. 30, n. 4, p. 712–727, abr. 2008.

[Boureau et al. 2010]BOUREAU, Y. et al. Learning Mid-Level Features for Recognition. In: *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR'10)*. [S.l.]: IEEE, 2010.

[Brown e Lowe 2002]BROWN, M.; LOWE, D. Invariant features from interest point groups. In: *In British Machine Vision Conference*. [S.l.: s.n.], 2002. p. 656–665.

[Cheng 1995]CHENG, Y. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE Computer Society, v. 17, n. 8, p. 790–799, Aug 1995.

[Clemons 2012]CLEMONS, J. *SIFT: Scale Invariant Feature Transform by David Lowe*. 2012. Acessado em: março de 2011. Disponível em: <www.eecs.umich.edu/ silvio/teaching/EECS598/lectures/lecture10_1.pdf>.

[Csurka et al. 2004]CSURKA, G. et al. Visual categorization with bags of keypoints. In: *In Workshop on Statistical Learning in Computer Vision, ECCV*. [S.l.: s.n.], 2004. p. 1–22.

[Duda, Hart e Stork 2001]DUDA, R. O.; HART, P. E.; STORK, D. G. *Pattern Classification (2nd Edition)*. 2. ed. [S.l.]: Wiley-Interscience, 2001.

[Garg, Vempati e Jawahar 2011]GARG, V.; VEMPATI, S.; JAWAHAR, C. V. Bag of visual words: A soft clustering based exposition. In: . [S.l.: s.n.], 2011.

[Gemert et al. 2010]GEMERT, J. C. van et al. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE Computer Society, Los Alamitos, CA, USA, v. 32, n. 7, p. 1271–1283, 2010. ISSN 0162-8828.

[Joachims 1998]JOACHIMS, T. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. 1998.

[Liu e Shah 2007]*Scene Modeling Using Co-Clustering*. 1–7 p.

[Lowe 2004]LOWE, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, Springer Netherlands, v. 60, p. 91–110, 2004. ISSN 0920-5691. Disponível em: <http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94>.

[MacQueen 1967]MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In: CAM, L. M. L.; NEYMAN, J. (Ed.). *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.]: University of California Press, 1967. v. 1, p. 281–297.

[Marszałek et al. 2007]MARSZAŁEK, M. et al. *Learning Object Representations for Visual Object Class Recognition*. oct 2007. Visual Recognition Challange workshop, in conjunction with ICCV.

[Papa e Rocha 2011]PAPA, J.; ROCHA, A. Image categorization through optimum path forest and visual words. In: *Intl. Conference on Image Processing*. [S.l.: s.n.], 2011. p. 3586–3589.

[Papa, Falcão e Suzuki 2009]PAPA, J. P.; FALCÃO, A. X.; SUZUKI, C. T. N. Supervised pattern classification based on optimum-path forest. *Int. J. Imaging Syst. Technol.*, John Wiley & Sons, Inc., New York, NY, USA, v. 19, p. 120–131, June 2009. ISSN 0899-9457. Disponível em: <http://dl.acm.org/citation.cfm?id=1541827.1541837>.

[Papa et al. 2011]PAPA, J. P. et al. Automatic classification of fish germ cells through optimum-path forest. In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. [S.l.]: IEEE Press, 2011. p. 5084–5087.

[Park e Jain 2008]PARK, S. P. U.; JAIN, A. K. Fingerprint verification using sift features. In: KUMAR, B. V.; PRABHAKAR, S.; ROSS, A. A. (Ed.). SPIE, 2008. v. 6944, n. 1, p. 69440K. Disponível em: <http://link.aip.org/link/?PSI/6944/69440K/1>.

[Philbin et al. 2008]PHILBIN, J. et al. Lost in quantization: Improving particular object retrieval in large scale image databases. In: *In CVPR*. [S.l.: s.n.], 2008.

[Queensland]QUEENSLAND, U. *COMP4702/COMP7703 - Machine Learning*. Acessado em: abril de 2012. Disponível em: <http://itee.uq.edu.au/ comp4702/lectures/k-means_bis_1.jpg>.

[Rocha, Cappabianco e Falcão 2009]ROCHA, L. M.; CAPPABIANCO, F. A. M.; FALCÃO, A. X. Data clustering as an optimum-path forest problem with applications in image analysis. *International Journal of Imaging Systems and Technology*, Wiley Periodicals, v. 19, n. 2, p. 50–68, 2009.

[Sande, Gevers e Snoek 2010]SANDE, K. van de; GEVERS, T.; SNOEK, C. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, Washington, DC, USA, v. 32, n. 9, p. 1582–1596, set. 2010.

[Shi e Malik 2000]SHI, J.; MALIK, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 8, p. 888–905, Aug 2000.

[Sinha 2010]SINHA, U. *SIFT: Scale Invariant Feature Transform*. 2010. Acessado em: maio de 2011. Disponível em: <http://www.aishack.in/2010/05/sift-scale-invariant-feature-transform/>.

[Sivic e Zisserman 2009]SIVIC, J.; ZISSERMAN, A. Efficient visual search of videos cast as text retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 31, n. 4, p. 591–606, 2009.

[Tong e Koller 2002]TONG, S.; KOLLER, D. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, JMLR.org, v. 2, p. 45–66, mar. 2002.

[Yang et al. 2007]YANG, L. et al. Discriminative cluster refinement: Improving object category recognition given limited training data. In: *CVPR*. [S.l.: s.n.], 2007.

Autorizo a reprodução xerográfica para fins de pesquisa.

São José do Rio Preto, 10 / 02 / 2013

Luis Alberto S. Afonso

Assinatura