VOCABULÁRIOS VISUAIS APLICADOS À DETECÇÃO DE EDIFÍCIOS EM FOTOGRAFIAS HISTÓRICAS

NATÁLIA COSSE BATISTA

VOCABULÁRIOS VISUAIS APLICADOS À DETECÇÃO DE EDIFÍCIOS EM FOTOGRAFIAS HISTÓRICAS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Arnaldo de Albuquerque Araújo

Belo Horizonte

31 de julho de 2009

Batista, Natália Cosse
B333v Vocabulários visuais aplicados à detecção de edifícios em fotografias históricas / Natália Cosse Batista. — Belo Horizonte, 2009
xxiv, 71 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Arnaldo de Albuquerque Araújo

1. Recuperação de Imagens com base no conteúdo. 2. Fotografias históricas. 3. Histograma de palavras visuais. 4. Detecção de edifícios. I. Orientador. II. Título.

CDU 519.6*84

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Vocabulários visuais aplicados à detecção de edifícios em fotografias históricas

NATALIA COSSE BATISTA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ARNALDO DE ALBUQUERQUE ARAÚJO - Orientador
Departamento de Ciência da Computação - UFMG

DR. EDUARDO ALVES DO VALLE JUNIOR
Bolsista de Pós-Doutorado-DCC-UFMG

PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG

PROF. GUILLERMO CÁMARA CHÁVEZ
Departamento de Computação - UFOP

Belo Horizonte, 31 de julho de 2009.

Dedico este trabalho ao meu marido, José Geraldo.

Agradecimentos

Agradeço primeiramente a Deus por permitir mais esta oportunidade de aprendizado e a Jesus, modelo a ser seguido.

Agradeço ao meu orientador, Arnaldo e "co-orientadores" pelas grandes contribuições para o início do meu aprendizado na vida acadêmica: Ana Paula e Eduardo.

Agradeço também aos professores Marcos André (UFMG), Guilhermo Chávez (UFOP), Mário Campos (UFMG) e David Menotti (UFOP) pelas valiosas sugestões. Aos demais professores e aos funcionários do DCC (Túlia, Sheila, Renata, Maristela, Alexandre) por sempre estarem dispostos a ajudar.

Agradeço a todos os colegas do NPDI: à Júlia pela amizade, ao Rodrigo pela colaboração em alguns trechos de código, ao David Flam pela força, à Sandra pelo ânimo, ao Eduardo Valle pelas várias dicas e explicações, ao Camillo pela colaboração no início do trabalho, Marcelo, Tiago, Pacheco, Thatyene e todos outros colegas pela agradável companhia. Em especial, à Ana Paula, que para mim é um exemplo de disciplina, pela imensa paciência e pelo encorajamento em momentos difíceis.

Agradeço aos meus pais, Cristina e José Agnaldo, que sempre investiram na minha educação e formação. Agradeço também a toda família pelo apoio, especialmente meus irmãos Tiago e Felipe e irmãs Bruna e Karina, avós (Vó Mary, Vô Luiz (*in memoriam*), Vó Zulmira, Vô Joaquim (*in memoriam*)), ao meu "irmão mais velho" Carlos e aos pais de meu marido, Vânia e Geraldo. Ao Mimi (*in memoriam*) pela docilidade e tão fiel companhia.

Agradeço aos amigos por compreenderem a minha ausência na maior parte do tempo (Grazi, Bruna, Vânia, Telminha, Natália, Fernanda...).

Agradeço aos colegas da UFOP (David, Sica, Alla, Regina, Fred, Ricardo Duarte, Lucília) pelas conversas e caronas nas quais aprendi muito também.

Agradeço ao apoio financeiro do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), fundamental durante este trabalho.

Principalmente agradeço ao meu marido, José Geraldo, que me acompanhou e apoiou em todos os momentos, pelo imenso amor e dedicação.

Resumo

Neste trabalho é proposta uma abordagem baseada na representação de histograma de palavras visuais para detectar imagens que contém edifícios na coleção de fotografias digitalizadas do Arquivo Público Mineiro (APM). As imagens utilizadas na avaliação do trabalho proposto provêm da digitalização dos originais, que é um procedimento efetivo para torná-los publicamente disponíveis e ao mesmo tempo preservá-los da degradação. A digitalização, porém, não resolve todos os problemas do acesso, que também envolve a necessidade de fornecer aos usuários meios para localizar os documentos desejados. A aplicação de técnicas de Recuperação de Imagens com Base no Conteúdo (RIBC) poderia ser de grande utilidade na descrição e indexação de grandes acervos históricos, que requer bastante tempo e esforço. Contudo, essas imagens tendem a ser mais desafiadoras para a aplicação de técnicas de RIBC do que as fotografias recentes, devido à perda de qualidade dos originais e diversos efeitos causados pela degradação. A técnica proposta utiliza um vocabulário visual como meio de construir um vetor de características para uma imagem e tem se mostrado eficaz além de robusta à oclusão e a variações devido à posição, escala, iluminação e várias outras transformações. Os resultados mostraram que, apesar da baixa qualidade das imagens, a abordagem é capaz de obter, para as imagens que contém edifícios, taxas médias de acerto próximas de 73%, indicando o potencial do método proposto para a tarefa de detecção.

Palavras-chave: Fotografias históricas; Detecção de edifícios; Vocabulários visuais; Histograma de palavras visuais; Classificação de imagens.

Abstract

In this work, we propose a technique based on a bag-of-keypoints representation to identify images containing buildings in the photographic collection of Arquivo Público Mineiro (APM), the Archives of the State of Minas Gerais. We evaluate the proposed work using digitized versions of the original. Digitization is an effective procedure to make collections publicly available, while preserving them from deterioration, but it does not solve, by itself, all the problem of access, which are also related to allowing the user to locate the desired items among thousands. The application of techniques of Image Retrieval Based on Content (RIBC) could be very useful in describing and indexing of large historical collections, which requires considerable time and effort. However, Archive images tend to be more challenging to implement RIBC techniques than recent photos, due to loss of quality of the original source and several effects caused by degradation. The proposed technique is based on a visual vocabulary which is used to build a feature vector for an image. This representation has been proved robust to occlusion and variations due to pose, scale, illumination and several transformations. Results show that, despite of the poor quality of the images, the bag-of-keypoints representation is able to provide detection rates around 73% for images containing buildings, indicating the suitability of the proposed method for the task of detection.

Keywords: Historical photographs; Image classification; Buildings detection; Bag-of-keypoints; Visual vocabulary.

Lista de Figuras

Lista de Tabelas

Lista de Siglas

APM - Arquivo Público Mineiro

DPI - *Dots per Inch*

PCA - *Principal Component Analysis*

PGM - *Portable Gray Map*

RBF - *Radial Basis Function*

RGB - *Red Green Blue*

RIBC - Recuperação de Imagens com Base no Conteúdo visual

SIFT - *Scale Invariant Feature Transform*

SVM - *Support Vector Machines*

TAEX - Transformada do Aguçamento EXtremo

Sumário

Capítulo 1

Introdução

A preservação de acervos de documentos históricos inclui, muitas vezes, a digitalização destes, que se apresenta como forma eficaz de viabilizar seu acesso público e equilibrar o severo compromisso entre conservação e acesso. A digitalização contribui para evitar o manuseio excessivo e a conseqüente aceleração do processo de degradação.

Devido à expansão do número de projetos de digitalização de acervos históricos e preservação digital da memória, um vasto número de imagens tem sido gerado (por exemplo, da Silva [2002], Wang et al. [2006] e Valle [2003]), representando centenas de gigabytes de informações.

A disponibilidade de grande quantidade de material digital de patrimônio cultural requer o desenvolvimento de novos métodos efetivos e de baixo custo para anotação e recuperação. A abordagem mais utilizada para esse fim é a chamada Recuperação de Imagens com Base no Conteúdo Visual (RIBC) e consiste basicamente em recuperar imagens similares a uma determinada especificação definida pelo usuário ou a um determinado padrão (por exemplo, esboço da forma, imagem exemplo, etc). Nessa abordagem são utilizadas técnicas de Processamento de Imagens e Visão Computacional que provêem meios de extrair informação útil de *pixels* e gerar uma descrição automática do conteúdo da imagem. A informação extraída pode ser, por exemplo, cor, textura, primitivas de baixo-nível (quinas, formas e relações espaciais) e informações de alto-nível (objetos, conteúdo da cena, descrição do assunto e até mesmo emoções associadas) [Chen et al., 2005].

Uma das vantagens da abordagem RIBC é a possibilidade de um processo de recuperação automático, em vez da abordagem tradicional de recuperação de informação baseada em palavras-chave, que requer a anotação prévia da base de imagens. A anotação manual de um grande acervo é um processo caro e demorado [Gevers & Smeulders, 2004; Wang et al., 2006], por isso fazem-se necessários sistemas que, além de

auxiliarem a anotação, permitam consultas de maneira eficaz e eficiente. Várias aplicações tem utilizado RIBC, tais como reconhecimento de impressões digitais, sistemas de informação de biodiversidade, bibliotecas digitais, prevenção de crimes, medicina e pesquisas históricas [Torres & Falcão, 2006].

1.1 Motivação

O Arquivo Público Mineiro[1] (APM) é uma superintendência da Secretaria Estadual de Cultura do Estado de Minas Gerais e contém documentos produzidos e acumulados por órgãos da Administração Pública de Minas Gerais e diversos arquivos privados. O APM recolhe e conserva importante patrimônio histórico e cultural, abrangendo os séculos XVIII, XIX e parte do século XX. Além de documentos manuscritos e impressos, reúne mapas, plantas, fotografias, gravuras, filmes, livros, folhetos e periódicos [Oliveira & de Albuquerque Araújo, 2004].

O acervo do APM é composto por aproximadamente 75.000 documentos, e desse total, as fotografias históricas representam uma parte significativa. A digitalização já foi realizada em cerca de 6.000 fotografias, que foram indexadas e disponibilizadas em um sistema de informação. Outras 14.000 encontram-se preparadas e organizadas para a digitalização futura. A Figura 1.1 mostra imagens de fotografias do acervo do APM, nas quais podem-se observar alguns exemplos do conteúdo e também da degradação visual.

Atualmente, no APM, a recuperação das imagens é realizada com base na informação textual, obtida pelo processo manual de indexação, que consiste na atribuição de termos de pesquisa aos itens do acervo. Esses termos podem ser palavras-chave controladas, provenientes de vocabulários controlados ou de tesauros (mais sofisticados que os vocabulários controlados pois permitem relacionar semanticamente os termos) [Valle, 2003]. A tarefa de descrição e indexação possui relação direta com as dimensões do acervo e com o número de conceitos anotados, podendo durar vários anos, pois em geral é uma tarefa realizada de forma manual, descrevendo-se um documento por vez [Kennedy et al., 2006]. Nesse contexto, ferramentas capazes de auxiliar no processo de anotação podem tornar-se valiosas [Oliveira & de Albuquerque Araújo, 2004].

Em uma análise realizada por Oliveira & de Albuquerque Araújo [2004] na base de imagens do APM, constatou-se que uma das necessidades da instituição é automação do processo de indexação, minimizando a interferência humana, que é, em parte, subjetiva.

[1]Endereço eletrônico: `http://www.siaapm.cultura.mg.gov.br/`. Acessado em 01 de junho de 2009.

Figura 1.1. Exemplos de documentos digitalizados da coleção de fotografias do APM.

Essa análise também incluiu entrevistas com os usuários do sistema de informação existente.

Em relação à recuperação de imagens, existe uma demanda pela identificação de objetos nas imagens do acervo (pontes, igrejas, vegetação, fachadas, placas, linhas de transmissão, linhas férreas, etc), que não é atendida pelo sistema de busca atual.

1.2 Objetivos

Dentre as inúmeras tarefas de RIBC que podem ser realizadas sobre o acervo fotográfico do APM, encontra-se a recuperação de imagens com fachadas de edifícios, casas, igrejas e outras construções. Neste trabalho, propõe-se uma metodologia para auxiliar na identificação automática de imagens dessa categoria específica, que pode ser usada na classificação de áreas entre urbanas ou rurais e na diferenciação entre ambientes externos e internos, colaborando com a indexação e recuperação dessas imagens, com base no conteúdo visual.

A recuperação de imagens pode ser realizada por meio do reconhecimento de categorias de objetos, no qual um detector especializado em uma classe de interesse é aplicado a uma dada imagem, determinando se um objeto dessa classe está presente ou não. O paradigma atual consiste em coletar manualmente um grande conjunto de bons exemplares da categoria do objeto desejado (conjunto de treino), treinar um classificador nesse conjunto e então avaliar o modelo obtido em novas imagens, possivelmente de natureza mais desafiadora [Fergus et al., 2005].

Dentro desse paradigma, o trabalho de Csurka et al. [2004] apresenta uma abordagem computacionalmente eficiente que tem sido bem sucedida para categorização de objetos e cenas. O método é denominado *bag-of-keypoints*, traduzido neste trabalho por "histograma de palavras visuais". O *bag-of-keypoints*[2] é uma analogia às representações *bag-of-words* usadas em recuperação de informação em textos, na qual cada documento é representado como um conjunto das palavras que ocorrem nesse documento [Nowak et al., 2006].

O método do histograma de palavras visuais e suas extensões (por exemplo Lazebnik et al. [2006], Perronnin [2008] e Marszalek & Schmid [2006]) têm se destacado recentemente atingindo bons resultados, devido à sua robustez e eficiência. Além disso, é capaz de lidar com uma grande variedade de objetos e cenas utilizando o mesmo algoritmo. Dadas todas essas características, neste trabalho optou-se por uma abordagem

[2]Alguns trabalhos utilizam a denominação *bag-of-features*, *bag-of-visual-words* ou *bag-of-words* referindo-se ao mesmo método.

baseada em histograma de palavras visuais para realizar a identificação de fotografias históricas contendo edifícios na base de imagens digitais do APM.

1.2.1 Objetivo geral

Desenvolver uma metodologia para identificação de fotografias contendo edifícios na base de imagens do APM, utilizando uma abordagem baseada em histogramas de palavras visuais.

1.2.2 Objetivos específicos

- Definir uma metodologia para elaborar o *ground-truth* de um subconjunto de imagens da base do APM;
- Aplicar uma abordagem baseada em histograma de palavras visuais no subconjunto de imagens;
- Classificar as imagens do subconjunto em relação à presença de edifícios, atribuindo-lhe uma das classes: edifícios, intermediárias ou não-edifícios; e
- Analisar a aplicação da abordagem para a tarefa proposta, tendo em vista as características próprias das fotografias históricas.

1.3 Organização da Dissertação

A dissertação está organizada em cinco capítulos e um glossário. No próximo capítulo (2) é apresentada uma revisão bibliográfica dos principais trabalhos encontrados na literatura relacionados à recuperação, classificação e detecção de edifícios e estruturas construídas pelo homem. Os principais conceitos envolvidos neste trabalho são explicados no Capítulo 3, que trata da abordagem utilizada. No Capítulo 4 os experimentos são descritos e seus resultados analisados. E o Capítulo 5 relata as conclusões, contribuições e trabalhos futuros propostos. O glossário contém uma breve explicação de alguns termos específicos utilizados ao longo do texto mas cuja definição detalhada foge ao escopo deste trabalho.

Capítulo 2

Trabalhos Relacionados

Neste capítulo são estudados os principais trabalhos encontrados na literatura relacionados com detecção, reconhecimento ou recuperação de imagens que contenham edifícios.

Em geral, reconhecimento está relacionado com a identificação de instâncias particulares de objetos. Detecção implica decidir se um membro de uma categoria visual está presente ou não em uma dada imagem (por exemplo a detecção de faces, carros ou pedestres). O problema da detecção de objetos também pode estar relacionado à determinação de sua localização na imagem, como definido em Agarwal et al. [2004]. Já recuperação de imagens refere-se ao processo de recuperar imagens similares a um padrão (por exemplo um esboço de forma, uma imagem exemplo, etc) ou a uma especificação definida por usuário, com base em descritores de características da imagem.

Os trabalhos relacionados a edifícios que envolvem esses problemas foram encontrados basicamente em três contextos: imagens aéreas e de satélites, fotografias ao nível do solo (*ground-level*) e vídeos.

2.1 Detecção de edifícios em imagens aéreas e de satélites

Em imagens aéreas e de satélites, a extração automática de edifícios e outros objetos feitos pelo homem é uma área de pesquisa que tem adquirido atenção significativa desde a última década [Karantzalos & Paragios, 2009; Mayer, 1999]. Várias aplicações necessitam de dados geométricos de cidades e outras localidades, provenientes de mapas bi- ou tridimensionais construídos com o auxílio dessas imagens. Tais aplicações incluem planejamento urbano, passeio (*tour*) virtual em cidades, planejamento da distribuição

de recursos de comunicação, gerenciamento de desastres naturais (enchentes, terremotos, etc), situações militares, localização de estatísticas geográficas (saúde, criminalidade, etc), dentre outras [Noronha & Nevatia, 2001; Saeedi & Zwick, 2008; Cord & Declercq, 2001].

A maioria das técnicas de detecção de edifícios em imagens aéreas e de satélite geram a hipótese da presença de telhados de edifícios na cena. Essas imagens apresentam propriedades bem diferentes de imagens de fotografias obtidas ao nível do solo, em que a detecção de telhados vistos de cima não é possível. Por ser um problema tratado de forma distinta, não estudaremos trabalhos que desenvolvem esta abordagem.

2.2 Recuperação, reconhecimento e detecção de edifícios em fotografias ao nível do solo

De um modo geral, pode-se classificar objetos de uma cena em dois tipos: objetos feitos pelo homem e objetos naturais. Objetos feitos pelo homem são caracterizados principalmente pela regularidade aparente e o relacionamento estrutural e espacial entre as características de seus componentes. Objetos naturais, como árvores, vegetação, rios, pedras e nuvens, podem coexistir com objetos feitos pelo homem, o que torna não-trivial a tarefa de detectar sua presença. Conforme definição de Iqbal & Aggarwall [1999]:

> Edifícios (e outras construções semelhantes) são objetos feitos pelo homem com bordas acentuadas e limites retos. A presença de um edifício na imagem gera um grande número de bordas, junções, linhas e grupos paralelos em comparação com uma imagem que predominantemente não contém edifícios. Essas estruturas podem ser generalizadas pela presença de quinas, janelas, portas, limites do edifício, etc. Essas características de nível intermediário exibem regularidade e proximidade, e são uma forte evidência da presença de estrutura na imagem. Ao passo que linhas retas extraídas de imagens de não-edifícios são geralmente distribuídas aleatoriamente. (Traduzido do original)

Nos trabalhos de Iqbal & Aggarwal [2002b] e Iqbal & Aggarwall [1999], regras do agrupamento perceptivo são aplicadas na recuperação por classificação de imagens contendo objetos de grandes dimensões feitos pelo homem, tais como edifícios, torres, pontes e outros objetos arquiteturais. Agrupamento perceptivo refere-se à habilidade

visual humana de extrair relações significantes de imagens a partir de características primitivas de baixo nível, sem nenhum conhecimento prévio sobre o conteúdo da imagem. Como mencionado no Capítulo 1, as características de baixo nível de uma imagem referem-se à cor, intensidade, contraste, textura, forma e localização espacial, sendo estritamente quantitativas. Já as características de alto-nível contêm informação semântica do conteúdo da imagem. O agrupamento de descrições de baixo nível da imagem provê uma estrutura de alto nível, que por sua vez pode ser combinada para originar estruturas de nível ainda mais alto. Esse processo pode ser repetido até que uma representação que contenha significado semântico seja atingida.

Segundo os autores, o sistema visual humano realiza uma construção hierárquica semelhante de características de baixo nível em representações de alto nível, utilizando conceitos como proximidade, similaridade, continuação, fechamento e simetria. Para detectar a presença de objetos feitos pelo homem a partir das características primitivas da imagem utilizando os princípios do agrupamento perceptivo, as seguintes características são extraídas hierarquicamente de uma imagem (Figura 2.1): segmentos retos, linhas compridas, coterminações, junções em "L", junções em "U", linhas paralelas, grupos paralelos, grupos paralelos significativos, grafo de coterminações e polígonos.

Nesses trabalhos, o espaço de características é particionado em três classes: estruturadas (imagens contendo estrutura significativa exibida por objetos feitos pelo homem), não-estruturadas (imagens não contendo objetos feito pelo homem) e intermediárias (imagens contendo uma mistura das duas classes). Na classificação das características extraídas e na recuperação das imagens utiliza-se o classificador dos k vizinhos mais próximos (*k-nearest neighbor*), que assinala cada imagem a uma das três classes.

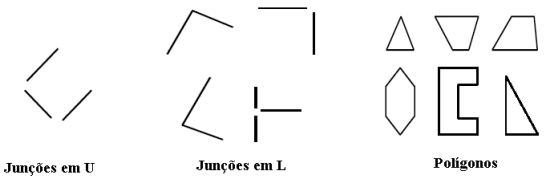

Figura 2.1. Exemplos de características extraídas de uma imagem com base no agrupamento perceptivo [Iqbal & Aggarwal, 2002b].

Uma das vantagens dessa abordagem é que a segmentação e a representação detalhada do objeto não são necessárias, ou seja, a decisão relativa à presença de objetos feitos pelo homem pode ser feita sem a necessidade de localizar e reconhecer um objeto específico, que requer maior conhecimento sobre as propriedades dos objetos. Outra vantagem é que a abordagem desenvolvida é invariante à reflexão, rotação e translação

dos dados da imagem [Iqbal & Aggarwal, 2001]. Entretanto, os autores afirmam que abordagens baseadas em agrupamento perceptivo são computacionalmente caras. Além disso, é necessário computar explicitamente as primitivas de baixo nível, o que requer que as imagens sejam relativamente livres de ruído [Kumar & Hebert, 2003].

Os autores também apresentam em Iqbal & Aggarwal [1999] um estudo comparativo do desempenho de sistemas de recuperação de imagens com base no conteúdo para localização de imagens que contêm edifícios ou objetos arquiteturais. Três descritores são comparados: histograma, textura e estrutura. Um estudo semelhante é realizado em Iqbal & Aggarwal [2002a]. De acordo com os autores, técnicas de análise de histograma e textura são incapazes de capturar descrições alto nível da cena, que relacionam diferentes características primitivas da imagem. Essas descrições são relativamente menos sensíveis a mudanças de iluminação em comparação com a análise de histograma e textura, de baixo-nível.

Ainda nesse trabalho, o histograma em escala de cinza normalizado foi computado, e, para análise de textura, utilizaram-se filtros de Gabor. A classificação das imagens foi feita com o classificador dos k vizinhos mais próximos. O histograma em escala de cinza é um descritor global que não permite relacionar diretamente as localizações espaciais na imagem. A textura trabalha com a análise da imagem em escalas locais, mas nas três classes há grande diversidade (por exemplo, imagens em close de uma superfície uniforme como um céu azul e imagens de vegetação possuem texturas que variam suavemente e rapidamente, respectivamente, embora ambas imagens pertençam à classe de não-edifícios).

Li & Shapiro [2002], em seu trabalho, partem da observação de que edifícios contém vários segmentos de linhas, freqüentemente horizontais e verticais, originadas de objetos diferentes e geralmente com cores diferentes ao seu redor. Para reconhecer e localizar objetos feitos pelo homem é utilizado o agrupamento de linha coerente (*consistent line cluster*), uma característica de nível intermediário local, que explora cor, orientação e características espaciais de segmentos de linha. Esse trabalho integra um sistema [Li et al., 2005] que reconhece uma variedade de objetos e conceitos em imagens e que pode ser usado para indexação automática e semi-automática de grandes bases de imagens.

O algoritmo realiza a detecção de bordas e linhas para extrair segmentos de linha da imagem. Para cada segmento de linha, são computados sua orientação e pares de cores (que geralmente estão presentes nas duas regiões que o segmento de linha faz fronteira). Os segmentos de linhas são então reunidos de acordo com os pares de cores, formando grupos de linhas coerentes por cor (*color-consistent line cluster*). Esses grupos são reagrupados com base na orientação, obtendo-se um conjunto de grupos de

linhas coerentes por orientação (*orientation-consistent line cluster*). Finalmente, os grupos obtidos são novamente reagrupados de acordo com suas posições na imagem (considerando segmentos próximos horizontalmente e verticalmente) para obter um conjunto final de grupos de linhas espacialmente coerentes (*spatially-consistent clustering*).

O conjunto final de grupos de linhas espacialmente coerentes é usado para detectar edifícios e outras estruturas em que predominam segmentos de linhas, por meio de dois critérios: as inter-relações entre os grupos (consideram as intersecções entre as extensões dos segmentos) e as intra-relações (consideram o número de linhas que se sobrepõem dentro do grupo). Para a recuperação de imagens foram construídos histogramas globais baseados nesses critérios e utilizado um classificador de árvore de decisão simples.

Conforme Kim et al. [2006], a desvantagem desse algoritmo é que seu funcionamento está restrito a edifícios altos e que contenham muitas bordas para cada grupo formado. Se o edifício for relativamente baixo e ocupar uma pequena porção da imagem, a taxa de detecção torna-se significativamente reduzida.

Outro trabalho que realiza o reconhecimento e localização de objetos feitos pelo homem é Kumar & Hebert [2003], que extrai características de blocos de 16×16 *pixels* da imagem. Para obter as informações de linhas retas e bordas das estruturas feitas pelo homem, a magnitude do gradiente e a orientação em cada *pixel* são determinados, e para cada bloco da imagem, é computado um histograma de orientações dos gradientes, ponderado pela magnitude do gradiente em cada *pixel*. Cada bloco é classificado individualmente como estruturado ou não-estruturado utilizando um modelo baseado em um campo causal multiescalar aleatório (*causal multiscale random field*). Como as características são computadas em múltiplas escalas, é possível capturar as dependências intra- e inter-escalares da vizinhança sem a necessidade de detecção explícita de bordas.

Zhou & Suter [2008] também utilizam essas mesmas características, derivadas de "orientogramas" (histogramas de orientações do gradiente em uma região ponderados pela magnitude dos gradientes) e imagens selecionadas da mesma base. As imagens são também divididas em blocos de 16×16 *pixels*, que são rotulados como edifícios ou não-edifícios. Os autores propõem um método para melhorar o desempenho da detecção e segmentação de edifícios baseado na classificação por processos Gaussianos.

Dado um conjunto de dados, a partir da transformação de Fourrier dos dados de treino, as características de cada dimensão no domínio da freqüência são analisadas para estimar a transformação de escala no domínio espacial. A escala dos dados tem o objetivo de minimizar as diferenças no espectro dos dados e torná-los isotrópicos,

ou seja, com propriedades homogêneas em cada dimensão. O espectro dos dados de treino é comparado com vários espectros dos *kernels* candidatos. A partir dessa comparação, o melhor *kernel* compatível é escolhido. Os resultados da classificação são comparados com os obtidos em [Kumar & Hebert, 2003] e relata-se uma melhor taxa de detecção e menos falsos positivos. Além disso, afirma-se que não é uma abordagem computacionalmente cara.

Yuan & Li [2006] apresenta uma abordagem para recuperar imagens de edifícios. Inicialmente, aplica-se a transformada de Hough no mapa de bordas da imagem (detectadas usando o algoritmo de Canny) para revelar a distribuição linear no domínio da transformada, que pode apresentar picos formados por pontos posicionados na mesma linha no mapa de bordas. Em seguida, o domínio da transformada de Hough é particionado em um número determinado de bandas e calcula-se o centróide dos picos em cada banda. Os centróides dos picos são usados para formar o vetor de características que descreve a forma de cada edifício na imagem. A similaridade entre duas imagens é avaliada pela disparidade dos vetores de características.

2.3 Detecção de edifícios em vídeos

Embora abordagens de detecção de edifícios em vídeos levem em consideração a estrutura do vídeo (por exemplo, informação temporal), as técnicas de detecção empregadas em quadros (*frames*) isoladamente podem ser aplicadas a imagens e serão analisadas.

O trabalho Hu et al. [2008] propõe uma abordagem para detectar e realizar o rastreamento (*tracking*) de edifícios de apartamentos para o desenvolvimento de um sistema de navegação baseada em vídeo. O objetivo, nesse caso é prover uma representação de realidade aumentada para informação de direção em seqüências de vídeo ao vivo, sendo aplicável também em veículos autônomos e robótica. O reconhecimento de edifícios é essencial para substituir objetos de destaque no mapa digital pela representação de realidade aumentada, por exemplo, sobrepor os nomes dos edifícios nas imagens desses edifícios em tempo real.

Nessa abordagem, uma cascata de classificadores fortes (construídos por técnicas de *boosting* para cada primitiva) é usada para detectar padrões representando apartamentos usando-se características de Haar. Baseado na localidade espacial dos padrões detectados, um agrupamento hierárquico aglomerativo é adotado para combiná-los em grupos que representam candidatos a apartamentos. Apesar dos autores relatarem resultados satisfatórios, pode ocorrer detecção incorreta de objetos em forma de barra como postes de sinalização em estradas, considerados errôneamente como parte de um

edifício. Oclusão severa causada por árvores, por exemplo, e ângulos de visão em que poucos padrões de apartamentos são observados podem também causar falha na detecção.

Outra abordagem para detecção de edifícios para navegação de robôs em ambientes urbanos é proposta por Trinh et al. [2007] e possibilita que o robô obtenha a altura e largura do edifício, sua cor e consequentemente o reconhecimento de um edifício específico [Trinh & Jo, 2006b]. A detecção de edifícios em imagens é realizada por meio da análise de suas propriedades geométricas e visuais.

Características tais como contraste de cor, direção e distribuição dos componentes principais dos edifícios (janelas, portas, colunas, regiões de parede) são usadas para refinar os segmentos de linha encontrados na imagem (essas características são detalhadas em Trinh & Jo [2006a]). Como os componentes principais no mundo real são geralmente paralelos, a projeção bidimensional de suas linhas determina um ponto de fuga dominante (*dominant vanishing point*). Segmentos de linha paralelos com um ponto de fuga em comum são agrupados criando uma rede de paralelogramos que representa a face de um edifício. Essa representação é usada para determinar se uma imagem contém ou não edifícios. Os componentes principais são formados pela união de paralelogramos vizinhos que possuem cores similares.

Observa-se nesse trabalho um grande número de parâmetros e limiares *ad hoc*, o que prejudica a utilização da técnica em outras bases de imagens. A base de imagens utilizada foi obtida a partir de fotografias de edifícios em duas grandes cidades (Ulsan e Zurique [Shao & Gool, 2003]), e contém imagens que em geral apresentam mais de um edifício, tipicamente com vários andares de apartamentos.

Kim et al. [2006] propõem um algoritmo para reconhecimento de edifícios para sistemas de navegação de veículos. Nesse algoritmo, as áreas da imagem que não contém edifícios – aquelas que contém estradas, árvores, veículos e outros objetos do fundo (*background*) são removidas por meio de uma máscara. A máscara é construída utilizando-se as características de pequenos blocos (por exemplo 10×10 *pixels*) nos quais a imagem é dividida. Os blocos são classificados com base na inclinação das bordas (verticais, horizontais, cruzadas ou variadas) e uma análise da vizinhança dos blocos determina se a área contém árvores e se é de fundo (por exemplo, blocos que pertencem à classe de inclinações variadas e estão conectados na direção vertical são rotulados como folhas/galhos).

Após remover esses objetos, é realizada uma busca por regiões de edifícios. Essa busca é determinada pelo uso de histograma de bordas nas direções horizontais, verticais e bordas cruzadas para cada bloco, seguido pela conexão dos blocos de classes relacionadas para determinar a área do edifício na imagem.

2.4 Considerações finais

Para o problema de recuperação, reconhecimento e detecção de edifícios observou-se que predominam nas técnicas a detecção de bordas e utilização de linhas. Uma das vantagens das abordagens que utilizam essas características é que a segmentação e a representação detalhada do objeto não são necessárias, ou seja, a decisão relativa à presença de objetos feitos pelo homem pode ser feita sem a necessidade de localizar e reconhecer um objeto específico. Entretanto, isso requer maior conhecimento sobre as propriedades dos objetos, por exemplo a aparência de componentes individuais dos objetos e suas relações espaciais, limitando a abordagem devido à pouca capacidade de generalização.

Até o limite de nosso conhecimento, todos os trabalhos encontrados na literatura utilizam bases de imagens de fotografias recentes, contendo em sua maioria edifícios modernos, com vários andares de apartamentos e consequentemente grande repetição de padrões lineares. Deve-se notar que essas bases possuem características distintas da base de fotografias do APM, cujos edifícios geralmente têm poucos andares e em que o desgaste, esmaecimento, estado de conservação, artefatos da digitalização, dentre outros, tornam a detecção de edifícios mais desafiadora.

As limitações dos métodos estudados motivaram a utilização de um método com maior capacidade de generalização, que mantivesse boa capacidade discriminativa mesmo com a grande variedade de instâncias visualmente diferentes de um tipo de objeto. Abordagens que expressam o estado da arte em RIBC são baseadas no reconhecimento de categorias de objetos [Chang et al., 2007], e são construídas com certo grau de robustez a variações da aparência das instâncias, tais como posição, iluminação, técnica de imageamento, etc. O método de histogramas de palavras visuais citado no Capítulo 1, no qual este trabalho se baseia, é um exemplo desse tipo de abordagem. Esse método consiste em modelar a distribuição de características de baixo-nível extraídas localmente de imagens desprezando-se suas localizações relativas e absolutas na imagem, sendo robusto a variações intra-classe.

O próximo capítulo apresenta a abordagem escolhida.

Capítulo 3

Metodologia

Neste capítulo são descritas as etapas da abordagem de histogramas de palavras visuais e os algoritmos utilizados na implementação do método.

3.1 Introdução

Basicamente, pode-se descrever a abordagem de histogramas de palavras visuais pelas seguintes etapas:

- Descrição de pontos característicos da imagem;
- Elaboração de um vocabulário de palavras visuais;
- Construção de um histograma de palavras visuais para cada imagem, que conta o número de descritores associados a cada palavra do vocabulário; e
- Aplicação de um classificador, tratando o histograma como vetor de características, para determinar qual categoria associar à imagem.

A idéia básica desse método é descrever uma imagem como uma coleção não-ordenada de características locais em torno de regiões de pontos de interesse, que são pontos da imagem que possuem grande quantidade de informação em termos de mudanças locais no sinal [Agarwal et al., 2004]. Vários métodos podem ser usados para amostrar os pontos de interesse de uma imagem, por exemplo densamente, aleatoriamente ou utilizando um detector de pontos de interesse. Um detector pode-se basear em critérios tais como contraste local, maximização ou minimização local de determinadas funções (Laplaciano, gradiente, etc) e limiarização sobre uma função de curvatura (Harris, Hessian, etc) [Valle et al., 2006]. Uma vez que o ponto é detectado, um descritor

deve ser gerado para descrever localmente esse ponto (geralmente apenas uma pequena região em volta do ponto é analisada).

Para uma representação compacta, um vocabulário visual é geralmente construído por meio do agrupamento (*clustering*) desses descritores. Cada grupo (*cluster*) de pontos de interesse é tratado como uma palavra visual no vocabulário, representada pelo centróide do grupo. Diferentemente do vocabulário textual de recuperação de informação em textos, o tamanho do vocabulário é determinado pelo número de grupos de pontos de interesse. Um vocabulário reduzido pode não ter boa capacidade de discriminação pois dois pontos de interesse podem ser associados ao mesmo grupo mesmo que eles não sejam similares um ao outro. Um grande vocabulário, por outro lado, é menos generalizável, menos complacente com ruído e causa sobrecarga de processamento [Jiang et al., 2007]. Tamanhos típicos de vocabulários visuais encontrados na literatura variam de 100 a 1000 palavras.

Por meio do mapeamento dos pontos de interesse em uma imagem ao vocabulário visual, pode-se descrever a imagem como um vetor de características de acordo com a presença ou contagem de cada palavra visual [Jiang et al., 2007]. Esse vetor de características é o *bag-of-keypoints* de uma imagem, ou histograma de palavras visuais (pois neste trabalho o vetor de características é computado pela contagem das palavras). Intuitivamente, espera-se que uma determinada categoria tenha algumas palavras visuais dominantes, por exemplo edifícios teriam uma dimensão dominante correspondendo à maior freqüência da palavra associada ao conceito "janela", ocorrendo o contrário com a palavra associada ao conceito "folha", que seria predominante em uma imagem de vegetação, como exemplificado na Figura 3.1 (uma comparação semelhante pode ser encontrada no trabalho de Sharma et al. [2008]).

Finalmente, um classificador pode ser aplicado aos vetores de características de determinado conjunto de imagens para determinar à qual categoria cada imagem pertence. A Figura 3.2 apresenta um esquema das etapas da abordagem.

De acordo com Nowak et al. [2006], as quatro principais decisões de implementação são: como amostrar pontos de interesse, como descrevê-los, como caracterizar as distribuições resultantes e como classificar as imagens com base no resultado. Várias abordagens foram sugeridas na literatura, com variações em cada etapa, como mostrado na Tabela 3.1. Essa tabela resume as decisões de implementação dos principais trabalhos encontrados na literatura, indicando os métodos utilizados nos trabalhos para amostrar e descrever pontos de interesse, para construir o vocabulário (o tamanho do vocabulário usado é indicado por k), para caracterizar e classificar a imagem.

As seções que seguem apresentam como a abordagem de histogramas de palavras visuais foi implementada neste trabalho.

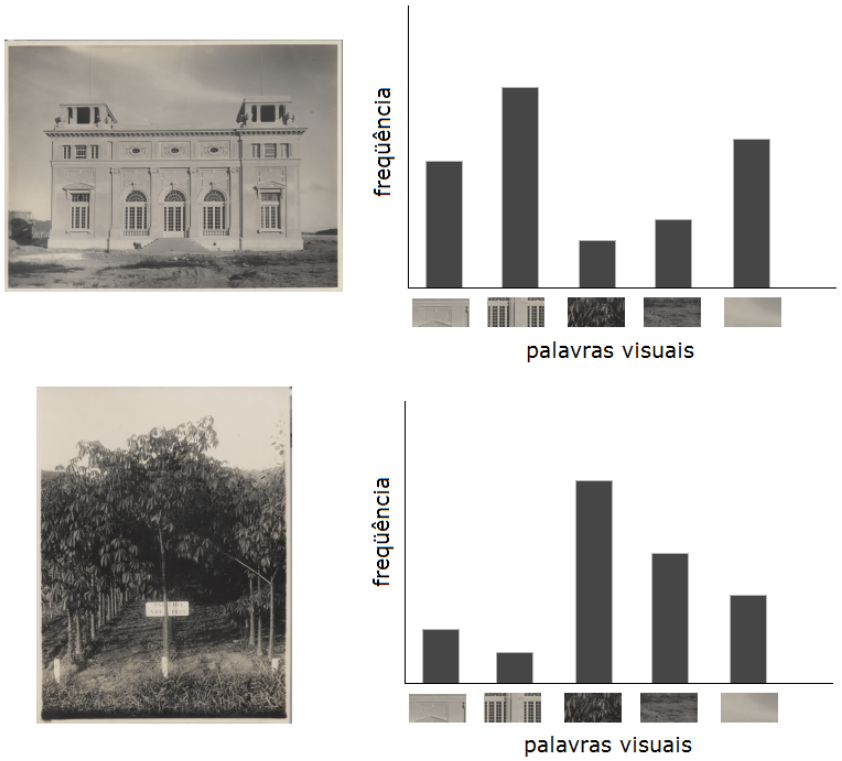

Figura 3.1. Visualização simplificada do histograma de palavras visuais.

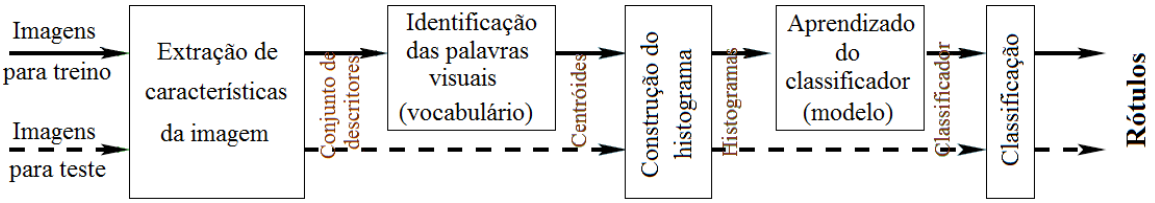

Figura 3.2. Esquema das etapas da abordagem de histogramas de palavras visuais. Figura traduzida e adaptada de Farquhar et al. [2005], página 2.

3.2 Estrutura do algoritmo

A Figura 3.3 mostra um esquema da estrutura do algoritmo para classificação das imagens do APM. Após o pré-processamento das imagens da base (explicado na Subseção 3.3), a detecção e descrição dos pontos de interesse é realizada utilizando o algoritmo SIFT (Subseção 3.4). Devido à alta dimensionalidade dos descritores SIFT, é utilizado o algoritmo PCA (Subseção 3.5) para reduzir o tamanho do vetor de descritores.

Os algoritmos são aplicados em imagens da base do APM, separadas em subconjuntos distintos para construir o vocabulário, treinar e testar o classificador. Utilizam-se os descritores dos pontos de interesse retirados das imagens do primeiro subconjunto

Trabalho	Amostragem de pontos	Descrição	Vocabulário	Caracterização	Classificação
Csurka et al. [2004]	Detector de Harris	SIFT	*k-means*, $k = 1000$	Histograma espacialmente ponderado	*Naive Bayes*, SVM com *kernel* linear
Farquhar et al. [2005]	Detector de Harris	SIFT	Modelos de Mistura de Gaussianas e *k-means*	Descritores SIFT reduzidos com PCA e PLS (*partial least squares*)	SVM com *kernel* linear
Jurie & Triggs [2005]	Amostragem densa em várias escalas	Método de similaridade ZNCC (*zero-mean normalized cross-correlation*) da intensidade dos *pixels*	Algoritmo de agrupamento dos autores, baseado em deslocamento pela média (*mean-shift*), $k = 600$	Histograma e arranjos binários	*Naive Bayes* e SVM com *kernel* linear
Nowak et al. [2006]	Amostragem aleatória, detector de Harris e Laplaciano de Gaussiana	SIFT e intensidades normalizadas dos pixels	*k-means*, $k = 1000$ e $k = 4000$	Histograma (contagem simples, binarização e informação mútua)	SVM com *kernel* linear e Gaussiano
Lazebnik et al. [2006]	Detector de Harris e grade regular	SIFT	*k-means*, $k = 1000$	Histograma de pirâmides espaciais	SVM com *kernel* de intersecção de histogramas
Marszalek & Schmid [2006]	Detectores de Harris-Laplace e Laplaciano	SIFT	*k-means*, $k = 1000$	Histograma espacialmente ponderado	SVM com *kernel* χ^2- Gaussiano
Jiang et al. [2007]	SIFT	SIFT	*k-means*, $k = 1000$	Histograma com ponderação suave	SVM com *kernel* χ^2- Gaussiano
Perronnin [2008]	Grade regular	SIFT e cor (média e desvio padrão)	Modelos de Mistura de Gaussianas	Conjunto de histogramas (um por classe)	Modelo de regressão logística organizada (SLR), SVM com *kernel* linear

Tabela 3.1. Síntese das principais abordagens que utilizam vocabulários visuais para categorização de imagens.

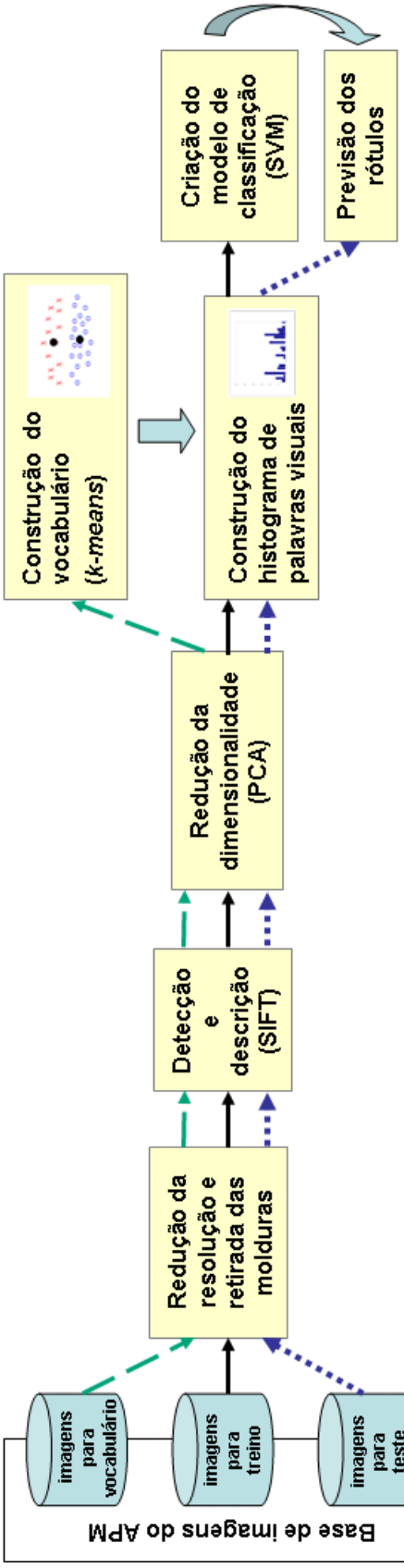

Figura 3.3. Principais etapas da abordagem de histograma de palavras visuais.

para construir o vocabulário, que é formado pelas palavras visuais resultantes do agrupamento desses descritores usando o algoritmo *k-means* (Subseção 3.6).

O histograma de palavras visuais dos subconjuntos de treino e teste são então calculados (Subseção 3.7). A classificação é realizada criando-se um modelo obtido das imagens de treino e aplicando esse modelo às imagens de teste, usando o algoritmo que implementa SVM (Subseção 3.8).

3.3 Pré-processamento

As imagens da base do APM encontram-se originalmente na resolução de 600 dpi (*dots per inch*) , no formato TIFF (*Tagged Image File Format*) e foram digitalizadas utilizando-se o sistema de cores RGB (*Red Green Blue*). Para redução da complexidade computacional, as imagens foram reduzidas a 72 dpi e armazenadas no formato PGM (*Portable Gray Map*), em tons de cinza.

Em seguida, foram retiradas manualmente as molduras das imagens, que incluem bordas decorativas ou falhas no processo de digitalização. As molduras são ruidosas e demonstraram afetar significativamente os resultados em tarefas de RIBC e indexação de imagens selecionadas desta mesma base [Lopes et al., 2008b], pois as características extraídas das molduras geram padrões não desejados. A Figura 3.4 mostra exemplos de fotografias com e sem molduras.

As imagens posicionadas fora da orientação normal foram também rotacionadas para facilitar a avaliação dos usuários, que será apresentada na Seção 4.1.

3.4 SIFT

A idéia do SIFT (Transformação de características invariantes à escala, do inglês *Scale-Invariant Feature Transform*) [Lowe, 1999] é detectar pontos de interesse em imagens, robustos à presença de transformações geométricas e variações de intensidade. O algoritmo então associa um descritor a cada ponto de interesse encontrado. Os descritores foram originalmente desenvolvidos para realizar o casamento de um objeto ou de uma cena em imagens diferentes, mesmo sob essas condições.

O SIFT é um dos métodos mais utilizados em abordagens de histogramas de palavras visuais e um dos que apresenta os melhores resultados em diversas tarefas [Nowak et al., 2006; Jiang et al., 2007]. Esse método transforma as informações de uma imagem em um conjunto de coordenadas relativas às características locais, razoavelmente invariantes à escala, mudanças de iluminação, rotação, perspectiva e ruído.

Figura 3.4. Imagens da base do APM. A primeira linha contém imagens com molduras; a segunda linha contém as mesmas imagens da primeira, porém com os pontos SIFT detectados em vermelho, inclusive nas molduras; e a terceira linha contém as imagens da primeira sem as molduras, que foram retiradas manualmente.

Cada ponto de interesse é associado a uma localização na imagem, uma escala e uma orientação.

O método SIFT consiste basicamente em quatro fases, que são resumidas a seguir. Mais detalhes sobre o algoritmo podem ser vistos nos trabalhos de Lowe [2004], Lowe [1999] e Cruz [2008].

1. *Detecção de pontos extremos em espaços multi-escala.*

   Nesta fase é realizada a detecção de locais que são invariantes a mudanças de escala da imagem, fazendo-se uma busca pelas características estáveis por várias escalas, usando uma função contínua Gaussiana. O espaço da escala é definido como uma função, $L(x, y, \sigma)$, obtida pelo resultado da convolução da imagem de entrada $I(x, y)$ com a função Gaussiana de escala variável $G(x, y, \sigma)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \tag{3.1}$$

onde $*$ é a operação de convolução em x e y e

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}. \tag{3.2}$$

Por questões de eficiência, Lowe [1999] propõe a detecção de pontos de interesse estáveis utilizando os extremos no espaço da escala obtidos com a convolução da função de diferença de Gaussianas na imagem, $D(x, y, \sigma)$, que pode ser computada pela diferença de duas escalas próximas separadas por um fator constante multiplicativo k:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma). \end{aligned} \tag{3.3}$$

$D(x, y, \sigma)$ representa a diferença entre imagens submetidas a um filtro Gaussiano em escalas σ e $k\sigma$. Este filtro consegue detectar variações de intensidade na imagem, tais como contornos. Como ilustrado na Figura 3.5, a imagem inicial é convolucionada incrementalmente para produzir imagens separadas por um fator constante k no espaço da escala (primeiro nível da pirâmide de escalas, à esquerda da figura). As imagens adjacentes são subtraídas para produzir as imagens das diferenças de Gaussianas (à direita da figura). Para gerar o próximo nível da pirâmide, a imagem da diferença de Gaussianas que possui o dobro do valor do σ inicial do nível anterior é escolhida e reduzida à metade de sua resolução usando interpolação bilinear. Esta será a primeira imagem do próximo nível, que obedecerá ao mesmo procedimento do nível anterior.

Para detectar um máximo ou mínimo local de $D(x, y, \sigma)$, cada ponto é comparado com seus oito vizinhos mais próximos na imagem corrente e nove vizinhos na escala acima e abaixo (Figura 3.6). O ponto é selecionado nessa se ele for maior ou menor que todos seus vizinhos, sendo considerado um potencial ponto de interesse.

2. *Localização de pontos de interesse.*

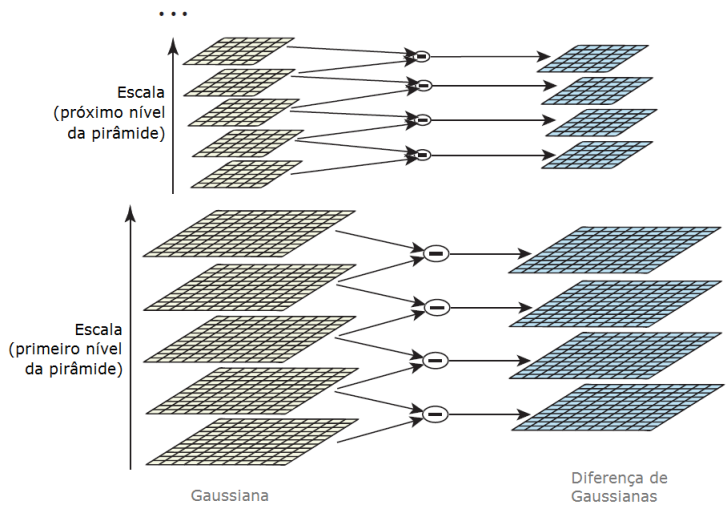

Figura 3.5. Abordagem eficiente para a construção de $D(x, y, \sigma)$. Imagem traduzida de Lowe [2004], página 6.

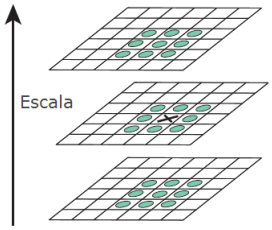

Figura 3.6. Detecção de extremos locais. Os pontos de interesse iniciais do SIFT consistem em extremos locais calculados a partir de filtros de diferença Gaussianos em múltiplas escalas. O *pixel* marcado com X é comparado com seus 26 vizinhos (marcados com círculos) em regiões 3×3 na escala corrente e nas escalas adjacentes. Imagem traduzida de Lowe [2004], página 7.

Para cada ponto de interesse candidato, determina-se a localização e a escala, sendo também realizada uma seleção dos pontos com base em medidas de sua estabilidade. Nesta etapa, pontos que possuem baixo contraste ou não estão próximos a uma quina são rejeitados.

A localização do extremo local é estimada por meio da utilização de uma função quadrática na região dos pontos amostrados, que realiza uma interpolação da posição. Para retirar os pontos que possuem baixo contraste, e portanto sensíveis a ruído, descartam-se os extremos cujo valor em relação à função D esteja baixo de um determinado limiar. A eliminação dos pontos de interesse próximos de apenas uma borda (considerados instáveis) é feita considerando-se o tamanho das curvaturas principais de $D(x, y, \sigma)$, pois esses pontos possuem uma curvatura principal grande ao longo de uma borda mas uma curvatura pequena na direção perpendicular.

3. *Cálculo da orientação.*

   Uma ou mais orientações são associadas a cada localização de ponto de interesse baseadas na direção do gradiente local para se obter a invariância quanto à rotação. Calcula-se para cada *pixel* da imagem na escala determinada a magnitude e a orientação do gradiente usando as diferenças de *pixels*.

   Um histograma das orientações dos *pixels* em uma região vizinha do ponto de interesse é construído, sendo que picos nesse histograma correspondem a direções dominantes para os gradientes locais. O maior pico do histograma e outros picos que correspondam a no mínimo 80% do valor do maior pico serão usados para criar um ponto de interesse com aquela orientação, ou seja, para locais de múltiplos picos de magnitude semelhante, são criados diferentes pontos de interesse na mesma localização, mas com diferentes orientações. Finalmente, uma parábola é usada para interpolar os três valores do histograma mais próximos do pico, de forma a se ter uma melhor precisão de sua posição.

4. *Descrição do ponto de interesse.*

   Esta etapa consiste em computar o descritor que represente as regiões relativas aos pontos de interesse. O descritor é criado computando-se as magnitudes e orientações dos gradientes que são amostrados ao redor da localização do ponto, na escala selecionada. Para alcançar a invariância à orientação, as coordenadas do descritor são rotacionadas em relação à orientação do ponto de interesse, calculada na etapa anterior.

   A Figura 3.7 ilustra o cálculo do descritor do ponto de interesse. Primeiramente, as magnitudes do gradiente e orientações são amostradas em volta da localização do ponto de interesse, usando a escala do ponto. Por questões de eficiência, os gradientes são pré-computados para todos os níveis da pirâmide (como representado pelas setas pequenas em cada localização amostrada no lado esquerdo da figura). Uma função Gaussiana com σ igual à metade da largura da janela do descritor é usada para associar um peso à magnitude do gradiente de cada ponto amostrado (como ilustrado no lado esquerdo da figura com a janela circular). O objetivo dessa janela Gaussiana é evitar mudanças repentinas no descritor com pequenas mudanças na posição da janela, e dar menos ênfase aos gradientes que estão distantes do centro do descritor.

   Uma vez feita a suavização dos gradientes, criam-se histogramas de orientação sobre regiões da amostragem. O descritor é formado por um vetor contendo as magnitudes de todas as orientações dos histogramas correspondentes aos tama-

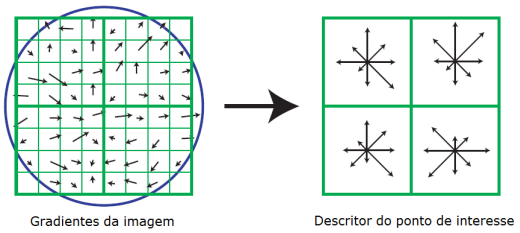

Figura 3.7. O descritor do ponto de interesse é criado computando-se a magnitude do gradiente e a orientação em cada ponto amostrado de uma região em volta do ponto de interesse. Os gradientes são ponderados por uma janela Gaussiana, indicada pelo círculo. As amostras são acumuladas em histogramas de orientação (8 direções) para cada sub-região, que no exemplo mostrado à direita tem tamanho 4×4. O tamanho das setas corresponde à soma das magnitudes dos gradientes próximos àquela direção. Imagem traduzida de Lowe [2004], página 15.

nhos das setas mostradas no lado direito da Figura 3.7. Essa figura mostra uma matriz 2×2 de histogramas de orientações, porém no algoritmo são utilizadas matrizes 4×4, resultando em um vetor de 128 elementos (4×4 histogramas de 8 orientações cada um) para cada ponto de interesse, pois mostraram os melhores resultados em experimentos realizados no trabalho de Lowe [2004].

O vetor gerado é normalizado para reduzir os efeitos de mudança de iluminação. Uma mudança de brilho na imagem na qual uma constante é adicionada a cada *pixel* não irá afetar os valores dos gradientes, pois eles são computados através da diferença de *pixels*, o que torna o descritor robusto a variações afins de iluminação. Variações não-lineares de iluminação (por exemplo devido à saturação da câmera ou mudanças de iluminação que afetem superfícies tridimensionais com orientações distintas) afetam a magnitude de alguns gradientes e em menor proporção suas orientações. Para reduzir o efeito das variações não-lineares, a magnitude de gradientes são limiarizadas e então normalizadas novamente. Níveis relevantes de robustez à distorções de forma também são atingidos devido à amostragem por regiões, o que permite que pequenos deslocamentos de posição dos gradientes ainda permitam a contribuição destes para os mesmos histogramas originais [Lowe, 2004; Marszalek & Schmid, 2006].

Ao final desse processo, são construídos vários descritores, cada um referente a um ponto de interesse, que capturam a estrutura espacial e a distribuição da orientação local de uma região que envolve o ponto de interesse. Quando se aplica o algoritmo SIFT em uma imagem, tem-se como resultado, portanto, um conjunto de descritores.

A Figura 3.8 mostra os pontos de interesse detectados pelo SIFT em algumas

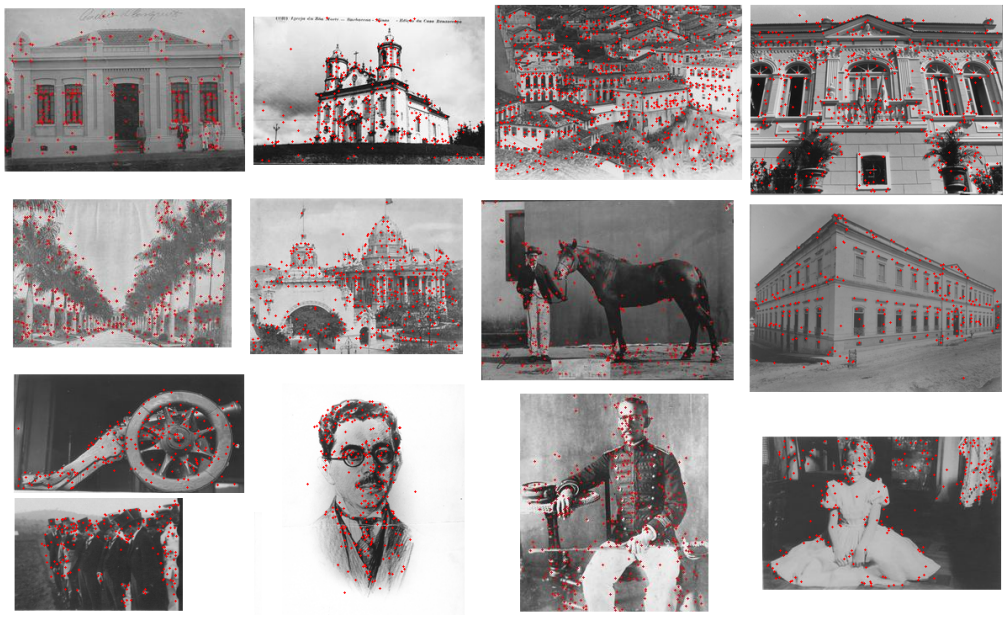

Figura 3.8. Imagens ao acervo do APM marcadas com pontos detectados pelo SIFT (cruzes vermelhas). O número médio de pontos SIFT detectados na base de imagens utilizada neste trabalho é de 385 por imagem.

imagens do acervo do APM. Neste trabalho foi utilizada a implementação do SIFT disponibilizada por David Lowe[1].

3.5 Algoritmo PCA

Para a construção do vocabulário visual é necessário realizar o agrupamento dos descritores dos pontos de interesse obtidos na etapa anterior. O algoritmo utilizado para isso é o *k-means*, que possui limitações quanto ao agrupamento de conjuntos de dados de alta dimensionalidade, por exemplo em relação ao problema dos mínimos locais e à dificuldade de amostrar um espaço de parâmetros muito grande, pois o método convencional pode gerar resultados bastante diferentes dependendo da amostra inicial [Ding et al., 2002]. Para melhorar o desempenho da próxima etapa de processamento, o método da Análise dos Componentes Principais (PCA, do inglês *Principal Components Analysis*) é utilizado para reduzir o vetor de descritores.

PCA é uma técnica estatística capaz de identificar padrões em dados e expressá-los de tal forma que suas diferenças e similaridades sejam realçadas, além de possibilitar a compressão dos dados pela redução do número de dimensões, sem grande perda de informação [Smith, 2002]. Esta técnica tem sido aplicada em diversas áreas tais como

[1] *Demo Software: SIFT Keypoint Detector*, disponível em `http://www.cs.ubc.ca/~lowe/keypoints`. Acessado em 01 de junho de 2009.

reconhecimento de faces, compressão de imagens e detecção de padrões em dados com alta dimensionalidade.

No PCA, as novas características são funções lineares das características originais. Dado um conjunto de pontos de dados, constrói-se um sub-espaço linear de dimensões menores que melhor explica a variação desses pontos a partir da média deles [Forsyth & Ponce, 2006]. De maneira simplificada, pode-se resumir em cinco os passos do algoritmo conforme Smith [2002]:

1. Subtrair dos dados de cada dimensão a sua média. Isso produz um conjunto de dados cuja média é zero. Seja um conjunto de dados $X = (x_1, x_2, x_3, \ldots, x_n)$ de uma determinada dimensão, com n elementos. Para cada dimensão, a fórmula da média é dada por:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.4}$$

2. Calcular a matriz de covariância. A covariância entre os conjuntos de dados de duas dimensões X e Y é dada por:

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n - 1} \tag{3.5}$$

   A matriz de covariância de um conjunto de dados de n dimensões é definida por:

$$C = (c_{X,Y}, c_{X,Y} = cov(X, Y))_{n \times n}, \tag{3.6}$$

   onde C é a matriz com n linhas e n colunas, $1 \leq X \leq n$ e $1 \leq Y \leq n$. Cada entrada da matriz é o resultado do cálculo da covariância entre duas dimensões distintas.

3. Calcular os autovetores e autovalores da matriz de covariância. Os autovetores da matriz de covariância permitem extrair linhas que caracterizam os dados. Os próximos passos envolvem transformar os dados de tal forma que serão expressos em termos dessas linhas.

4. Escolher os componentes e formar um vetor de características. Uma vez que os autovetores são calculados, o próximo passo é ordená-los do maior para o

menor, obtendo os componentes em ordem de significância. Os componentes menos significativos podem ser ignorados com perda de informação relativamente pequena, e o conjunto de dados finais terá consequentemente menos dimensões que o original. Precisamente, se originalmente existem n dimensões nos dados, serão calculados n autovetores e n autovalores. Como os p primeiros autovetores serão escolhidos, o conjunto final de dados terá apenas p dimensões.

A matriz de características é construída a partir dos autovetores que são mantidos, formando uma matriz tendo esses vetores como colunas:

$$MatrizDeCaracterísticas = [autovetor_1 \quad autovetor_2 \quad \dots \quad autovetor_p] \tag{3.7}$$

5. Derivar o novo conjunto de dados. Uma vez determinada a matriz de características, multiplica-se a transposta da matriz à esquerda do conjunto original de dados, também transposto:

$$DadosFinais = MatrizDeCaracterísticas^T \quad \times \quad DadosAjustados^T, \tag{3.8}$$

onde $MatrizDeCaracterísticas^T$ é a matriz dos autovetores com as colunas transpostas de forma que os autovetores são agora linhas dessa matriz, com o autovetor mais significativo na primeira linha, e $DadosAjustados^T$ são os dados ajustados pela média transpostos, isto é, os dados de cada ponto estão em uma coluna, com cada linha representando um dimensão distinta. $DadosAjustados^T$ é o conjunto final de dados, com dados de cada ponto em uma coluna, e dimensões ao longo das linhas. $DadosFinais$ é o conjunto original de dados em termos dos vetores escolhidos.

Este método é utilizado para reduzir os descritores SIFT obtidos na etapa anterior de 128 para 30 dimensões, que em testes preliminares demonstrou bom compromisso entre informações mantidas e tamanho do vetor. Os descritores das imagens para vocabulário são usados para construir a matriz de características e um vetor de médias das dimensões, que serão necessários para realizar o ajuste e a redução dos descritores das imagens de treino e teste.

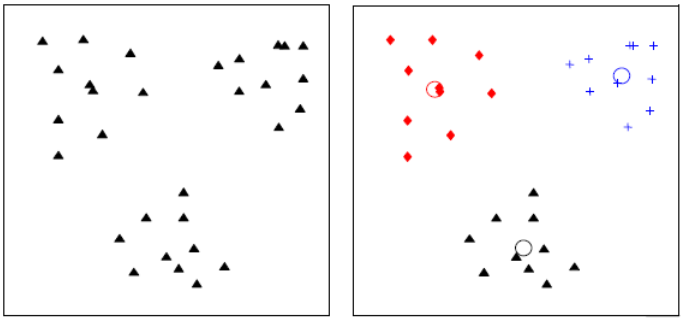

Figura 3.9. Exemplo de agrupamento no espaço bidimensional. (a) Os pontos (ou padrões) são posicionados no espaço com um triângulo. (b) Agrupamento obtido pelo *k-means*, onde com $k = 3$ e pontos que pertencem ao mesmo grupo possuem o mesmo símbolo. Os centróides de cada grupo estão marcados com círculos. Figura extraída de Jain [2009], página 7.

3.6 Algoritmo *k-means*

Métodos de classificação não-supervisionados são utilizados para agrupar dados dos quais não se conhece o rótulo, sendo conhecidos também como métodos de agrupamento (*clustering*), em que os rótulos associados a cada categoria são obtidos apenas por meio dos dados. Neste trabalho o agrupamento é realizado com o algoritmo *k-means*, que é um método simples e com complexidade computacional moderada [Jain et al., 1999].

O *k-means* particiona um conjunto de pontos entre k subconjuntos disjuntos de forma a minimizar a distância intra-classe e maximizar a distância inter-classe. A Figura 3.9 mostra um exemplo simples de agrupamento com 3 grupos.

O algoritmo *k-means* pode ser resumido nos seguintes passos:

1. Escolher aleatoriamente k pontos que serão os centros de grupos iniciais, denominados centróides.

2. Para cada ponto do conjunto de treinamento, associá-lo ao centro do grupo mais próximo, usando uma medida de distância entre o ponto e o centróide dos grupos (por exemplo, distância Euclidiana, cosseno do ângulo entre os pontos, etc).

3. Recalcular os centróides dos grupos usando os grupos correntes formados.

4. Se o critério de convergência não for atingido, voltar para o passo 2. Critérios de convergência típicos são:

   - nenhuma (ou mínima) reassociação de pontos a novos centróides de grupos; ou

Figura 3.10. Parte de um determinado grupo do vocabulário visual das imagens do APM contendo 128 regiões de 13×13 *pixels*. O centro de cada região é um ponto de interesse pertencente ao grupo. O número médio de pontos de interesse utilizados para construir o vocabulário é 37210 (utiliza-se uma fração do total, no caso 40%, por razões de eficiência).

- decréscimo mínimo na soma dos erros quadrados, que é a soma das distâncias entre cada ponto e o centróide mais próximo; ou
- o número de iterações atingidas.

Os centróides dos grupos resultantes do agrupamento realizado pelo *k-means* forma o vocabulário dos pontos de interesse representativos. Utiliza-se a distância Euclidiana e o critério de convergência é determinado pelo valor mínimo da soma dos erros quadrados. A Figura 3.10 mostra um exemplo com as palavras visuais do vocabulário formado a partir do agrupamento de pontos de interesse de um conjunto de imagens do APM. Nota-se que as palavras visuais desse exemplo estão relacionadas a padrões de janelas de edifícios.

3.7 Construção do histograma de palavras visuais

Para a construção do histograma de palavras visuais de uma imagem, é realizada a associação dos descritores dos pontos de interesse encontrados nessa imagem ao conjunto de grupos obtidos pelo *k-means*. Para essa associação é considerada a distância Euclidiana no espaço de características, resultando em uma contagem simples dos pontos de interesse de uma imagem mais próximos a cada palavra do vocabulário visual.

Seja uma imagem que contém um conjunto de descritores de pontos de interesse D e um histograma de k palavras visuais, que será representado pelo arranjo H.

Para computar o histograma de palavras visuais, são realizados os seguintes passos:

1. O arranjo H é iniciado com zero em todas as suas k posições.
2. Para cada um dos descritores do conjunto D:

- Calcular o arranjo D_i, com as distâncias Euclidianas do descritor i a todos os centróides dos k grupos.
- Determinar a menor distância do arranjo D_i e adicionar uma unidade ao elemento de H que corresponde ao grupo mais próximo.

3. Normalizar o histograma H, de tal forma que a soma de todos os elementos seja 1. Esse passo possibilita que cada elemento do histograma represente a proporção dos descritores na imagem que contém determinada palavra visual, e não seu número absoluto, que pode apresentar grandes variações dentro de uma mesma classe de imagens.

Os histogramas de palavras visuais gerados para as imagens serão utilizados como entrada para a próximo algoritmo, o do classificador, que os tratará como vetores de características.

3.8 Algoritmo de SVM

Neste trabalho escolheu-se para a classificação das imagens as Máquinas de Vetores de Suporte (SVM, do inglês *Support Vector Machines*). O SVM é uma técnica de Aprendizado de Máquina, que vem sendo utilizada em diversas tarefas de reconhecimentos de padrões, obtendo resultados superiores aos alcançados por outras técnicas de aprendizado em várias aplicações (categorização de textos, análise de imagens e bioinformática) [Lorena & de Carvalho, 2007; Jiang et al., 2007; Csurka et al., 2004]. De acordo com Lorena & de Carvalho [2007], SVM apresenta como vantagens:

- Boa capacidade de generalização, que é medida por sua eficiência na classificação de dados que não pertençam ao conjunto utilizado em seu treinamento;
- Robustez diante de dados de grande dimensão; e
- O uso de funções *kernel* na não-linearização do SVM, permite representar espaços abstratos de forma eficiente pois seu cálculo é mais simples que o de uma função de mapeamento.

SVM está inserido dentro dos métodos de classificação supervisionada, que é um problema que pode ser definido como:

> Dado um conjunto de exemplos rotulados na forma $(x_i; y_i)$, em que x_i representa um exemplo e y_i denota o seu rótulo, deve-se produzir um classifi-

> cador, também denominado modelo, preditor ou hipótese, capaz de predizer precisamente o rótulo de novos dados. Esse processo de indução de um classificador a partir de uma amostra de dados é denominado treinamento. O classificador obtido também pode ser visto como uma função f, a qual recebe um dado x e fornece uma predição y [Lorena & de Carvalho, 2007].

Resumidamente, pode-se dizer que o objetivo do SVM é a construção de hiperplanos que separam os exemplos positivos e negativos de uma classe com a maior margem possível, que é uma medida de confiança da previsão do classificador. As próximas subseções (3.8.1 e 3.8.2) introduzem em mais detalhes as idéias nas quais o algoritmo de SVM é baseado. Os conceitos e definições foram retirados do trabalho de Lorena & de Carvalho [2007].

3.8.1 SVM linear

Seja T um conjunto de treinamento com n dados $x_i \in X$ e seus respectivos rótulos $y_i \in Y$, em que X constitui o espaço dos dados e $Y = \{-1, 1\}$. T é linearmente separável se é possível separar os dados das classes $+1$ e -1 por um hiperplano.

A equação de um hiperplano que divide o espaço dos dados X em duas regiões é dada por:

$$f(x) = \mathbf{w} \cdot \mathbf{x} + b = 0, \tag{3.9}$$

em que $\mathbf{w} \cdot \mathbf{x}$ é o produto escalar entre os vetores $\mathbf{w}$ e $\mathbf{x}$, $\mathbf{w} \in X$ é o vetor normal ao hiperplano descrito e $\frac{b}{||\mathbf{w}||}$ corresponde à distância do hiperplano em relação à origem, com $b \in \mathbb{R}$ (Figura 3.11).

Os dados que se encontram mais próximos ou sobre os hiperplanos são denominados vetores de suporte e são considerados os dados mais informativos do conjunto de treinamento, pois somente eles participam na determinação da equação do hiperplano separador. Como resultado final, obtém-se uma função linear otimizada com a melhor capacidade de generalização, que representa o hiperplano separador o qual divide os dados com a maior margem.

3.8.2 SVM não-linear

Em situações reais, é difícil encontrar aplicações cujos dados sejam linearmente separáveis devido a fatores como a presença de ruídos e discrepâncias (*outliers*) nos dados

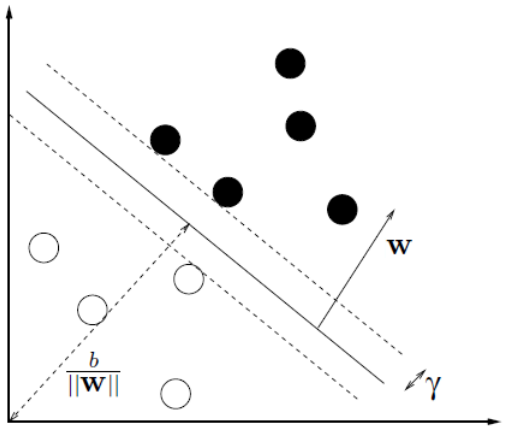

Figura 3.11. Hiperplano separador (representado pela linha cheia) das classes de dados (cada classe é representada por um símbolo distinto). O tamanho da margem é indicado por γ. Os outros elementos são discutidos no texto. Esta figura foi adaptada de Fradkin & Muchnik [2005], página 3.

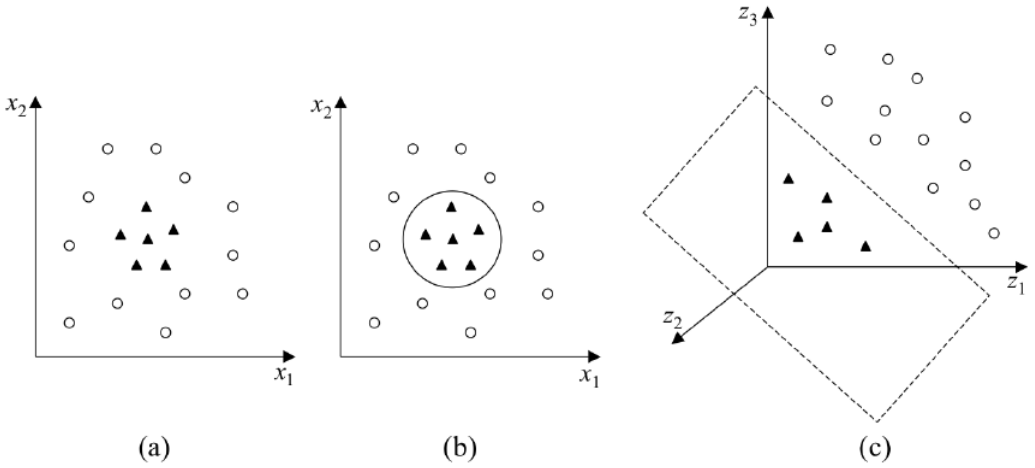

Figura 3.12. (a) Conjunto de dados não-linear. (b) Fronteira curva no espaço de entradas para a separação das classes. (c) Fronteira linear no espaço de características. Figura extraída de Lorena & de Carvalho [2007], página 18.

ou à própria natureza do problema, que pode ser não-linear. A Figura 3.12 (a,b) mostra um exemplo de dados desse tipo.

Nesse caso, o conjunto de treinamento é mapeado de seu espaço original para um novo espaço de maior dimensão, denominado espaço de características. Seja $\Phi : X \rightarrow \Im$ um mapeamento, em que X é o espaço original e $\Im$ denota o espaço de características. A escolha apropriada de Φ permite que o conjunto de treinamento mapeado em $\Im$ possa ser separado por um SVM linear, como mostrado no exemplo da Figura 3.12 (c).

A função Φ não precisa ser conhecida explicitamente se for utilizada uma função K no treinamento [Burges, 1998]. Essa função é denominada *kernel*, definida por:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j). \tag{3.10}$$

A função K recebe dois pontos $\mathbf{x}_i$ e $\mathbf{x}_j$ do espaço original e computa o produto escalar desses dados no espaço de características. Os principais *kernels* utilizados no SVM são apresentados na Tabela 3.2.

Tipo de *kernel*	Função correspondente	Comentários
Polinomial	$({\mathbf{x}_i}^T \cdot \mathbf{x}_j + 1)^p$	A potência p deve ser especificada pelo usuário.
Gaussiano	$\exp(-\frac{1}{2\sigma^2}\|\|\mathbf{x}_i - \mathbf{x}_j\|\|^2)$	A amplitude σ^2 é especificada pelo usuário.
Sigmoidal	$\tanh(\beta_0 \mathbf{x}_i \cdot \mathbf{x}_j + \beta_1)$	Utilizado somente para alguns valores de β_0 e β_1.

Tabela 3.2. Sumário dos principais *kernels* utilizados no SVM. Tabela extraída de Lorena & de Carvalho [2003].

A escolha do kernel e dos parâmetros considerados tem efeito no desempenho do classificador obtido, sendo uma limitação do método a sensibilidade à escolha dos valores dos parâmetros. Embora existam vários *kernels* de propósito geral, não é claro qual é o mais eficaz para abordagens de histogramas de palavras visuais no contexto da classificação visual. Entretanto, na maioria dos trabalhos existentes o *kernel* escolhido é o linear ou o Gaussiano (também conhecido por RBF, do inglês *Radial Basis Function*) [Jiang et al., 2007; Hsu et al., 2009]. Neste trabalho é utilizado o *kernel* Gaussiano pois produziu melhores resultados em testes preliminares.

O SVM é um classificador binário, ou seja, o número de classes utilizado deve ser igual a dois. Para aplicá-la a problemas multiclasses, existem abordagens tais como a decomposição "um-contra-todos" (*one-versus-all*) e "todos-contra-todos" (*all-versus-all*). No primeiro caso, sendo k o número de classes, produz-se k classificadores, cada um separando uma classe i das $k-1$ restantes. No segundo caso, são produzidos classificadores para separação de cada classe i de outra j, em que $i, j = 1, \cdots, k \quad e \quad i \neq j$. Neste trabalho é utilizada uma abordagem "um-contra-todos".

A implementação do SVM utilizada neste trabalho é o LIBSVM [Chang & Lin, 2001], uma biblioteca para Matlab[2] que pode ser utilizada em problemas de classificação multiclasses.

[2] `http://www.mathworks.com/products/matlab/`, acessado em 01 de junho de 2009.

3.9 Considerações finais

Como visto neste capítulo, várias abordagens que utilizam vocabulários visuais foram propostas na literatura. Todas elas apresentam basicamente as mesmas etapas (escolha e descrição de regiões de interesse, caracterização e classificação da imagem) para solucionar o problema da categorização visual, embora com implementações distintas. Entre as limitações do método de histogramas de palavras visuais, pode-se citar o fato de que não leva em consideração informações sobre a distribuição espacial das características e é incapaz de capturar forma ou de segmentar um objeto do fundo (*background*) da imagem.

Também foram apresentadas as decisões de implementação das etapas do método empregado neste trabalho, detalhando-se os algoritmos utilizados. No próximo capítulo são apresentados os resultados dos experimentos, que consistem na avaliação do método de histogramas de palavras visuais na base de imagens históricas do APM.

Capítulo 4

Resultados experimentais

Neste capítulo são descritos os experimentos realizados e é mostrada uma análise dos resultados. A parte experimental consistiu basicamente de três fases:

1. Construção do *ground-truth*: uma metodologia baseada na consulta a usuários foi desenvolvida para anotar as imagens da base utilizada nos experimentos.

2. Experimentos preliminares: realizados com o objetivo de testar a influência de alterações na abordagem de histograma de palavras visuais nos resultados da classificação.

3. Experimentos finais: com base nos experimentos preliminares, foi escolhida a melhor configuração para ser aplicada em todo o conjunto de imagens selecionadas da base do APM.

Nas três fases experimentais foram utilizadas imagens de um conjunto selecionado aleatoriamente da coleção de fotografias digitalizadas do APM. Esse conjunto foi composto por 600 imagens, que correspondem a aproximadamente 10% do total de imagens digitalizadas do acervo e será referenciado como base experimental.

As seções seguintes detalham cada fase experimental e as métricas utilizadas para avaliação dos resultados.

4.1 Construção do *ground-truth*

Um dos aspectos mais importantes que afetam a classificação semântica é a qualidade da anotação [Chang et al., 2007]. Hanbury [2008] aponta como o melhor método para

criação de um *ground-truth*[1] para avaliação de um algoritmo a criação de um vocabulário de palavras-chaves textuais e, então, a anotação manual das imagens utilizando essas palavras. Entretanto, esse método está sujeito à inconsistência entre anotadores e, por consumir muito tempo, exige um vocabulário textual pequeno.

Para obter a anotação textual das imagens escolhidas foi estabelecida uma metodologia que obtém o rótulo de cada uma das imagens a partir da opinião de 3 usuários voluntários, que foram orientados sobre o contexto do problema. Com base no trabalho de Iqbal & Aggarwal [2002b], as categorias identificadas para o problema da detecção de edifícios foram: edifícios, não-edifícios e intermediárias.

No contexto de categorização visual, considera-se como *ground-truth* o conjunto de imagens corretamente anotadas com um texto descrevendo cada imagem, por exemplo, os objetos que aparecem nessa imagem, seu conteúdo semântico ou sua categoria, como no caso deste trabalho. O *ground-truth* permite utilizar métodos de aprendizado supervisionado e também avaliar a eficácia da classificação, possibilitando a comparação do desempenho do sistema contra o de um classificador humano. Esperou-se obter um *ground-truth* menos sujeito a erros em relação a uma anotação de apenas um usuário em cada imagem, pois é realizada uma votação que associa à imagem o rótulo escolhido pela maioria.

A elaboração da base anotada foi dividida em quatro etapas:

1. Seleção de um conjunto de imagens para os experimentos;
2. Construção de um sistema Web para anotação;
3. Avaliação das imagens pelos usuários; e
4. Computação dos resultados por votação.

Para a primeira etapa, as imagens da base experimental foram divididas em 6 grupos de 100 imagens cada (G1, G2, G3, G4, G5 e G6). Quando necessário, foram realizados ajustes na orientação das imagens que estavam rotacionadas. Foram substituídas as imagens que continham algum conteúdo potencialmente objecionável (por exemplo, cadáveres de pessoas).

Na segunda etapa foi construída uma interface Web na linguagem PHP para a anotação das imagens[2].

[1] *Ground-truth* é também conhecido como gabarito, verdade terreno ou padrão de ouro.

[2] A interface para anotação das imagens encontra-se hospedada nos endereços: `http://www.npdi.dcc.ufmg.br/anotacao/G1/`; `http://www.npdi.dcc.ufmg.br/anotacao/G2/`; `http://www.npdi.dcc.ufmg.br/anotacao/G3/`; `http://www.npdi.dcc.ufmg.br/anotacao/G4/`; `http://www.npdi.dcc.ufmg.br/anotacao/G5/`; e `http://www.npdi.dcc.ufmg.br/anotacao/G6/`.

A interface do sistema mostra uma imagem por vez e as três opções de classes, uma das quais o usuário deve marcar conforme sua opinião em relação à classificação da imagem (Figura 4.1):

- Edifícios: se o usuário julgar que aparece alguma fachada de prédio, casa, ou outra construção civil na maior parte da fotografia;
- Não-edifícios: se o usuário julgar que não se pode considerar que a fotografia contém fachadas de prédios, casas ou construções civis; e
- Intermediárias: se o usuário julgar que a fotografia contém alguma fachada de prédio, casa ou outra construção civil mas são pouco relevantes para a fotografia, por exemplo, esses objetos aparecem ocupando pequena porção da imagem (menos da metade) ou apresentam oclusão severa.

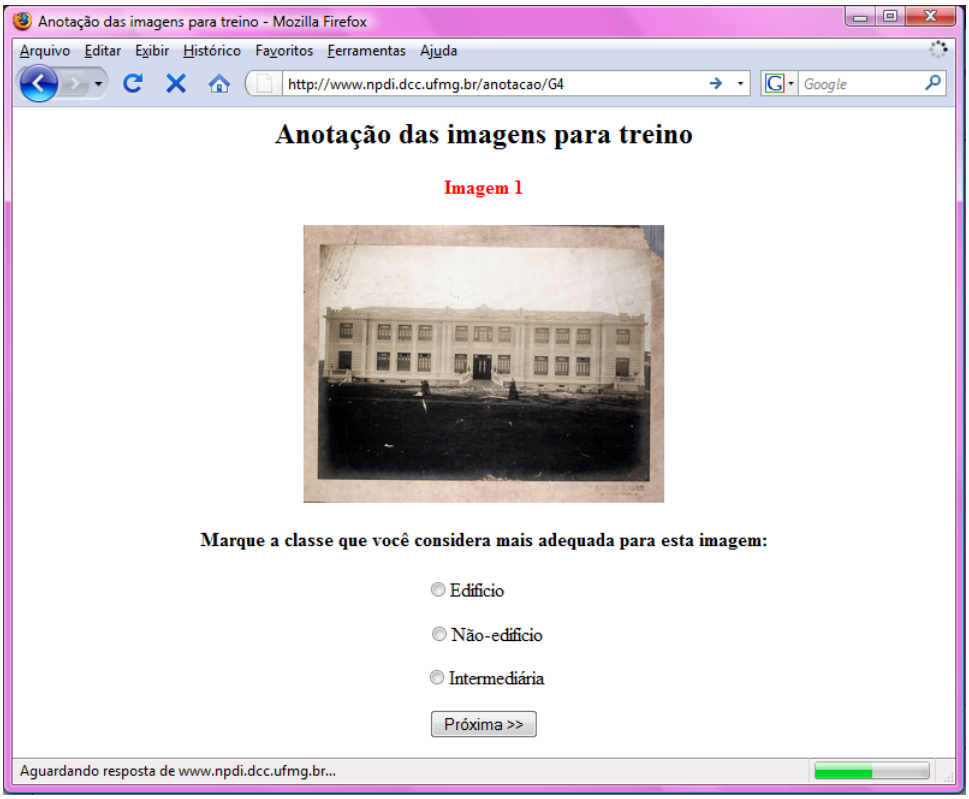

Figura 4.1. Instantâneo da interface do sistema Web para anotação das imagens pelos usuários.

A terceira etapa consistiu em convidar 18 usuários para realizar a anotação manual, sendo que cada usuário avaliou um dos grupos. Dessa forma, cada imagem de um grupo foi anotada por 3 usuários distintos.

Após a anotação dos usuários, na quarta etapa a classe a qual a imagem pertence foi definida como aquela escolhida pelo maior número de usuários, semelhante a uma votação. Por exemplo, se uma determinada imagem foi anotada como sendo da classe edifícios pelo primeiro usuário, da classe edifícios pelo segundo usuário e da classe intermediárias pelo terceiro usuário, então considerou-se que a imagem pertence à classe edifícios, escolhida pela maioria. Nos casos em que uma imagem foi anotada com 3 classes distintas (aconteceu em 3% das imagens), um quarto usuário foi chamado para avaliar, resolvendo o empate. Como resultado foram obtidas para cada classe:

- Edifício: 145 imagens;
- Não-edifício: 367 imagens; e
- Intermediária: 88 imagens.

4.2 Métricas de avaliação dos resultados

As métricas utilizadas nos experimentos são apresentadas a seguir:

- *Taxa de acerto*

  A taxa de acerto é a porcentagem de todas imagens classificadas corretamente, por exemplo a porcentagem de acerto das imagens de uma classe em relação ao número total de imagens dessa classe.

- *Acurácia*

  A acurácia da classificação, neste trabalho, é considerada como a média não-ponderada das taxas de acerto de cada classe. Essa métrica é melhor aplicada a conjuntos de imagens desbalanceados do que a métrica mais tradicional da porcentagem de todas imagens classificadas corretamente, que é induzida pelas classes mais representativas [Nowak et al., 2006].

- *Matriz de confusão*

  A matriz de confusão sumariza os resultados da classificação em duas dimensões, onde as linhas são os rótulos do *ground-truth* e as colunas indicam o rótulo atribuído pelo classificador. Cada elemento da matriz indica o número de imagens anotadas como a classe relativa à sua linha e classificada como a classe relativa à sua coluna. O número de predições corretas de cada classe encontram-se na diagonal principal da matriz. Todos os outros elementos correspondem ao número de imagens classificadas incorretamente.

Os resultados dos experimentos foram obtidos por um dos métodos:

- Teste de estimativa simples: divide o conjunto de imagens em dois subconjuntos mutuamente excludentes, um para treinamento e outro para teste.
- Validação cruzada (*cross-validation*): na validação cruzada com f dobras, o conjunto de imagens é dividido em f subconjuntos ou dobras mutuamente excludentes de tamanhos aproximadamente iguais. Em seguida um subconjunto é testado utilizando o classificador treinado com os $f - 1$ subconjuntos restantes. Neste caso, a acurácia da validação cruzada é dada pela média das acurácias obtidas das f execuções.

Em experimentos executados repetidas vezes, calculou-se o intervalo de confiança da acurácia da classificação (ou das taxas de acerto de uma classe separadamente), que permite comparar quantitativamente resultados experimentais de métodos distintos e verificar se a diferença entre eles é significativa. O intervalo de confiança é dado por limites probabilísticos, quais sejam c_1 e c_2, tal que existe uma grande probabilidade (nível de confiança) da média da população estar no intervalo (c_1, c_2) [Jain, 1991].

4.3 Experimentos preliminares

Como descrito no Capítulo 3, a abordagem de histograma de pontos de interesse apresenta diversas variações na literatura. Vários fatores podem afetar o desempenho dessa abordagem, como por exemplo o tamanho do vocabulário visual, o método para computar as palavras visuais de uma imagem, a função *kernel* utilizada no aprendizado supervisionado, dentre outros [Jiang et al., 2007]. O objetivo dos experimentos preliminares é testar algumas variações propostas e outras que foram bem sucedidas em trabalhos com imagens de bases recentes em imagens da base do APM, para se chegar a uma configuração experimental final a ser aplicada em toda a base experimental.

A abordagem usada nos experimentos, quando não especificado, utiliza:

- O detector e descritor de pontos de interesse SIFT;
- A redução dos descritores pelo PCA para 30 dimensões (40% dos descritores do conjunto de treino são usados para determinar a matriz de características do PCA);
- O vocabulário visual determinado pelo *k-means*;
- O histograma de palavras visuais construído por contagem simples; e

- O classificador SVM com *kernel* Gaussiano.

As próximas subseções contém uma descrição e análise dos experimentos preliminares realizados.

4.3.1 Processamento com realce de imagens

O objetivo deste experimento foi verificar se a utilização de métodos de filtragem e realce tradicionais nas imagens implicaria melhores resultados na classificação. Para tal, foram selecionados 8 métodos de realce que, em uma análise preliminar visual, demonstraram melhorar a qualidade das imagens. Alguns dos métodos aplicados às imagens também fizeram com que um maior número de pontos de interesse fossem amostrados, o que pode significar mais informação discriminativa para a classificação conforme Jiang et al. [2007]. Os métodos utilizados são descritos a seguir:

1. Deconvolução cega

   A restauração cega de imagens (*blind image restoration*) é o processo de estimar a imagem verdadeira e o processo de borramento (*blurring*) a partir das características da imagem degradada, utilizando informações parciais sobre o sistema de imageamento [Kundur & Hatzinakos, 1996]. A deconvolução cega (*blind deconvolution*) tem o objetivo de restaurar a imagem quando não há informações sobre a sua distorção (borramento e ruído). O algoritmo utilizado é baseado em métodos de Máxima Verossimilhança para derivar os filtros de restauração pela estimativa do operador de distorção. A estimativa inicial fornecida ao algoritmo foi uma função Gaussiana com $\sigma = 0,5$ e o número de iterações utilizadas foi 10.

2. Deconvolução cega pelo método de Lucy-Richardson

   Método com o mesmo objetivo do item anterior (1), porém com implementação mais eficiente que utiliza técnicas de otimização [Matlab, 2009]. A deconvolução cega baseada no algoritmo de Lucy-Richardson é utilizada para reconstruir imagens degradadas por altos níveis de ruído [Fish et al., 1995].

3. Filtro Gaussiano seguido de realce do constrate

   O filtro Gaussiano é um filtro de suavização (passa-baixas) usado para borramento e redução do ruído, resultando em uma imagem mais uniforme e contornos menos nítidos. Conforme Gonzalez & Woods [2003], o borramento é utilizado em pré-processamento para remoção de pequenos detalhes de uma imagem antes da

extração de objetos (grandes), e conexão de pequenas descontinuidades em linhas e curvas. Utilizou-se um filtro Gaussiano 3×3 com $\sigma = 0,5$. Em seguida, a imagem foi submetida a um filtro de realce do contraste (Item 8).

4. Alargamento de contraste seguido do filtro da mediana

   Uma função de transformação realiza mapeamento dos níveis de cinza de forma a produzir uma imagem de maior contraste do que a original. Ocorre o escurecimento de 1% dos níveis de cinza próximos do preto e o clareamento daqueles próximos do branco na imagem original. Os outros valores ajustados pela função de transformação no intervalo $[0, 1]$.

   Em seguida, o filtro da mediana 3×3 é aplicado na imagem, com o objetivo reduzir o ruído preservando a agudeza das bordas. O filtro da mediana consiste em substituir o nível de cinza de cada *pixel* pela mediana dos níveis de cinza na vizinhança daquele *pixel*. A mediana m de um conjunto de valores é tal que metade dos valores no conjunto são menores do que m e metade dos valores são maiores do que m (por exemplo, em uma vizinhança 3×3 a mediana é o quinto maior valor) [Gonzalez & Woods, 2003].

5. Filtro Laplaciano seguido do alargamento de contraste

   Filtro que aproxima o operador de derivada Laplaciano bidimensional. Este filtro detecta transições de cor, sendo propício para detectar contornos na imagem. Um imagem intermediária é obtida subtraindo-se a imagem original pela imagem processada pelo filtro Laplaciano (após uma limiarização que alterou os *pixels* com valores menores que 50 para 0, eliminando parte das bordas ruidosas). A imagem final é o resultado da imagem intermediária submetida ao alargamento de contraste.

6. Filtro de Prewitt seguido de filtro de Sobel

   Os filtros de Prewitt e de Sobel baseiam-se na magnitude do gradiente (primeira derivada) para acentuar as bordas horizontais. Os valores de gradiente menores do que o limiar 40 são transformados no valor mínimo preto (0). Aos *pixels* para os quais os valores de gradiente não obedecessem a esse critério foram atribuídos seus próprios valores originais na imagem. A imagem resultante é obtida pela combinação dessas imagens com a imagem original para mostrar as bordas horizontais realçadas.

7. Filtro da transformada do aguçamento extremo (TAEX)

O objetivo deste filtro é recuperar as bordas danificadas. Para uma janela $w \times w$ da imagem, o valor do operador extremo é escolhido entre dois níveis de cinza extremos dentro da janela. Aquele que tiver o valor mais próximo ao nível de cinza do *pixel* central da janela será utilizado (ou o próprio valor do *pixel*, se este se encontrar exatamente entre os dois valores extremos). Utilizou-se uma janela de tamanho 3×3.

8. Filtro de realce do contraste

   Filtro 3×3 criado pelo negativo do filtro Laplaciano com parâmetro $\alpha = 0,2$, que destaca os contornos da imagem.

O experimento consistiu na seleção aleatória de 300 imagens da base experimental e a validação cruzada com 3 dobras, utilizando os algoritmos implementados no pacote de Processamento de Imagens do *Matlab* (exceto TAEX, implementado pela autora). Para cada dobra, foi realizada uma varredura no espaço dos parâmetros com o objetivo de obter a melhor taxa de acerto. Utilizou-se o tamanho de vocabulário $k = 100$.

As Figuras 4.3 e 4.4 mostram a aplicação dos algoritmos de realce em imagens utilizados no experimento. Os resultados são exibidos na Figura 4.2. Observa-se no gráfico desta figura que não houve melhora significativa do resultado da classificação com a aplicação de realce nas imagens, o que demonstrou que os descritores SIFT sofreram pouca influência das alterações realizadas, já que apresentam certa robustez a mudanças de iluminação e ruído.

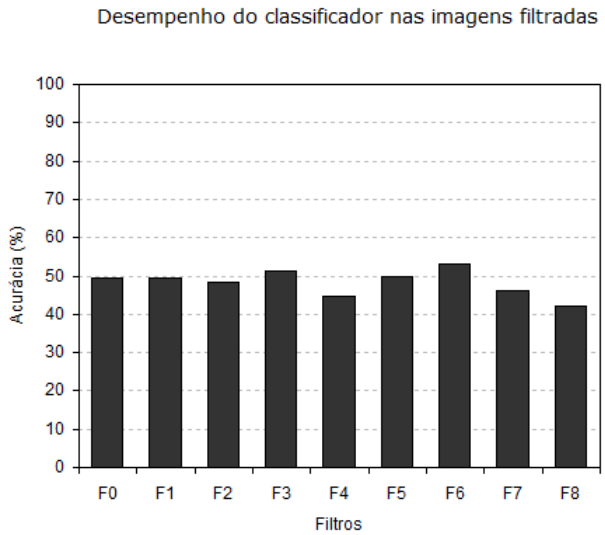

Figura 4.2. Gráfico da acurácia da classificação para vários filtros testados, com tamanho de vocabulário $k = 100$. F0 representa o teste com imagens sem filtros e F1 a F10 representam os filtros enumerados no texto.

Figura 4.3. Exemplos de imagens processadas com os filtros testados. F0 corresponde à imagem original sem pré-processamento e F1 a F8 representam os filtros enumerados no texto (a figura continua na próxima página).

Figura 4.4. Continuação da Figura 4.3, da página anterior. Exemplos de imagens processadas com os filtros testados. F0 corresponde à imagem original sem pré-processamento e F1 a F8 representam os filtros enumerados no texto.

4.3.2 Escolha do tamanho do vocabulário visual

Neste teste é analisado como o tamanho do vocabulário k afeta o desempenho da classificação. Optou-se por testar valores de k de 100 a 1.500, que abrange a faixa de tamanhos típicos de vocabulários visuais encontrados na literatura.

Foram testadas 380 imagens selecionadas aleatoriamente da base experimental, divididas em 180 para treino (60 de cada classe) e 200 para teste. Os resultados deste experimento podem ser vistos no gráfico da Figura 4.5, em que são computados a melhor acurácia obtida em cada teste. Nota-se que não existe uma tendência clara no relacionamento entre o número de palavras visuais e a acurácia da classificação (por exemplo, o aumento no número de palavras não implica no crescimento da acurácia), sendo que os valores de acurácia mantiveram-se por volta de 60%.

Como conclusão deste teste, observou-se que o agrupamento realizado está sujeito a flutuações aleatórias do *k-means*, pois não há um comportamento previsível nos resultados. Isso é devido ao fato que diferentes partições iniciais podem resultar em um agrupamento final distinto, pois o *k-means* converge apenas para um mínimo local

da função de critério, devido à natureza gulosa do algoritmo [Jain et al., 1999]. Além disso, o *k-means* não é um método eficaz para agrupar dados de alta dimensionalidade [Ding et al., 2002].

De acordo com Jurie & Triggs [2005], *k-means* é um método que funciona melhor em imagens contendo predominantemente regiões de textura uniforme. Regiões de interesse que são mais informativas para a classificação tendem a ter freqüências intermediárias. Regiões mais freqüentes tipicamente contém estruturas genéricas como bordas que ocorrem, em geral, em todas as classes, fornecendo pouca informação discriminativa. Por outro lado, regiões menos freqüentes são mais discriminativas, mas ocorrem tão raramente em uma imagem, que pouco influenciam no desempenho. Ainda conforme os autores, esses efeitos são provavelmente maiores em ambientes feitos pelo homem, onde são comuns grandes regiões uniformes pouco informativas.

Escolheu-se $k = 100$ para a realização dos experimentos finais, já que produziu resultados semelhantes a valores maiores de k e também devido ao fato que um número menor de palavras visuais possibilita uma redução no tempo de processamento, principalmente do algoritmo *k-means*.

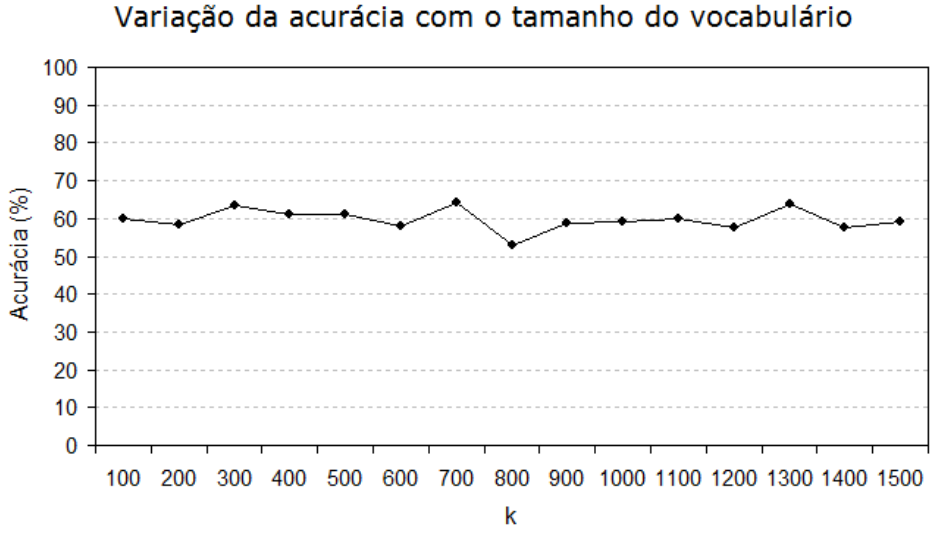

Figura 4.5. Gráfico da acurácia da classificação das imagens de teste para vários tamanhos de vocabulário. Foram utilizadas imagens das classes edifícios, não-edifícios e intermediárias para treino e teste.

4.3.3 Construção do vocabulário por meio de agrupamento separado

Este experimento consiste em construir o vocabulário de cada classe separado e então juntá-los em um vocabulário único, pela concatenação dos centróides. De acordo com

Farquhar et al. [2005], o agrupamento dos descritores por categorias separadas preserva informações discriminativas perdidas durante o agrupamento das categorias juntas. Essa idéia é ilustrada na Figura 4.6, em que o agrupamento separado levou a um melhor posicionamento dos centróides no exemplo mostrado.

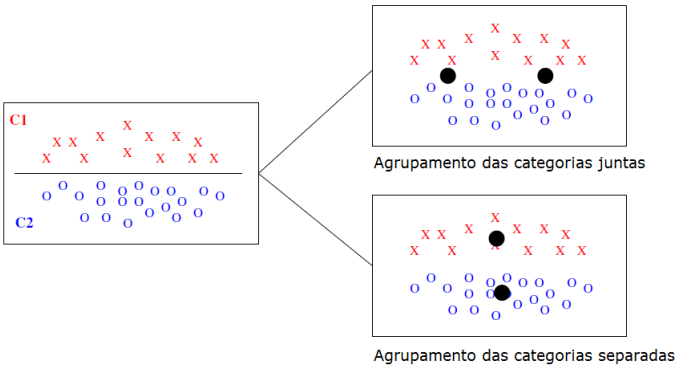

Figura 4.6. Agrupamento de categorias hipotéticas de dados. Os círculos cheios representam os centróides do grupo, que apresentam posições distintas conforme o tipo de agrupamento. Figura traduzida de Farquhar et al. [2005].

As imagens utilizadas para construir o vocabulário foram obtidas de recortes manuais realizados a partir de 200 imagens selecionadas aleatoriamente da base do APM, mas diferentes das imagens da base experimental. Partes das imagens que continham apenas edifícios e partes que continham apenas outros elementos (como carros, pessoas, céu, água, montanha, etc) foram recortadas, gerando 80 novas imagens. Desses recortes, 34 são de edifícios e 46 de outros objetos (Figura 4.7). Dessa forma garante-se que o vocabulário formado por uma categoria não contém elementos de outra.

Os descritores obtidos das imagens de recortes de edifícios foram agrupados utilizando *k-means* com $k = 50$ e o mesmo foi feito com as imagens de recortes de não-edifícios. Os centróides dos grupos obtidos para as duas categorias foram combinados para formar o vocabulário único de tamanho igual a 100.

As imagens utilizadas para os experimentos foram as mesmas da Subseção 4.3.1 e o resultado foi computado pela mediana do melhor resultado de cada uma das 5 dobras utilizadas na validação cruzada. Na Figura 4.10 podem ser comparados os resultados deste experimento com o experimento de vocabulário construído por agrupamento das categorias juntas, não sendo observada diferença significativa.

Como a granularidade da composição do descritor (que descreve uma pequena região em torno de um ponto de interesse) é menor do que o recorte feito nas imagens, basicamente os pontos no espaço dos descritores continuam sendo os mesmos que os das imagens originais. Se, por exemplo, os descritores de duas regiões semanticamente

distintas da imagem forem agrupados como uma mesma palavra visual e no agrupamento separado acontecer desses descritores serem associados artificialmente a duas palavras visuais distintas, conseqüentemente a construção do histograma de palavras visuais é não é melhorada. As duas palavras visuais produzidas são praticamente as mesmas em relação à proximidade, então a associação dos descritores das imagens de teste a essas palavras é problemática, sendo divididos entre as palavras mesmo sendo equivalentes.

4.3.4 Divisão da imagem em regiões

A representação de histogramas de palavras visuais não leva em conta as relações espaciais entre as características [Marszalek & Schmid, 2006], o que motivou a realização deste experimento. Para incorporar de maneira simples as relações espaciais, cada imagem foi dividida em regiões iguais e o histograma de uma região foi computado separadamente. O vetor de características dessa imagem foi composto pela concatenação dos histogramas das regiões.

O número de regiões escolhidas foi 4 para o primeiro teste e 9 para o segundo. Como ilustrado pela Figura 4.8, edifícios tendem a aparecer na fotografia geralmente na parte central e inferior. A idéia era que os histogramas dessas regiões indicassem maior concentração de palavras visuais relacionadas com edifícios.

O experimento foi realizado com 380 imagens selecionadas aleatoriamente da base experimental, sendo 180 para treino (60 para cada classe) e 200 para teste. Do total de descritores das imagens de treino, 50% foram selecionados aleatoriamente para realizar a redução com PCA. Para construir o vocabulário, foram utilizados 40% dos descritores também ds imagens de treino, sorteados aleatoriamente.

Os resultados deste experimento são mostrados no gráfico da Figura 4.9, em que foram registradas, para cada k, a melhor acurácia obtida. O aumento do número de regiões fez com que o desempenho da classificação fosse reduzido. Isso pode ser explicado pelo fato de que uma região menor da imagem contém menos pontos de interesse detectados e conseqüentemente menos descritores para caracterizar o conteúdo daquela região. Além disso, o aumento da dimensionalidade do vetor de características resulta em um desempenho mais pobre do classificador, problema conhecido na literatura como "maldição da dimensionalidade", pois, dentre outros fatores, a adição de mais características pode aumentar o ruído, e portanto o erro.

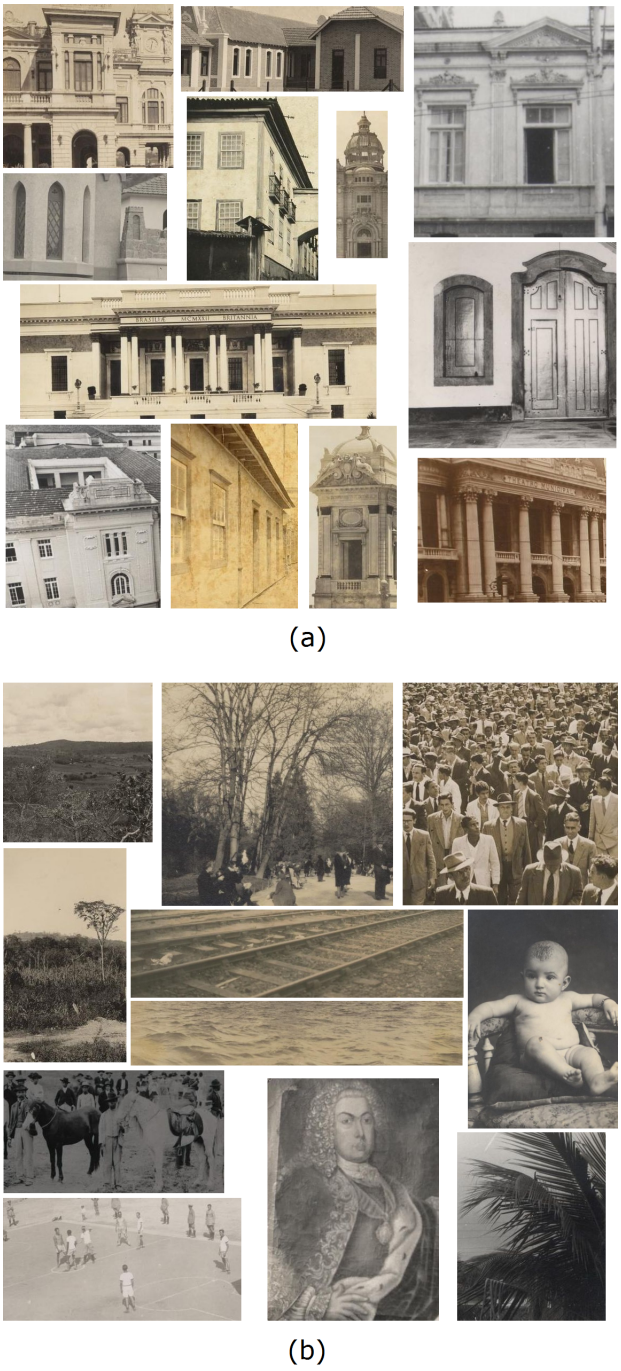

Figura 4.7. Recortes de imagens da base do APM:(a) Edifícios; e (b) Não-edifícios.

4.3.5 Construção do histograma de palavras visuais por meio de *soft-weighting*

Dentre os fatores que podem afetar o desempenho de métodos baseados em histogramas de palavras visuais estão os pesos atribuídos às palavras do histograma, conforme

Figura 4.8. Divisão da imagem em regiões para computar os histogramas de palavras visuais separadamente.

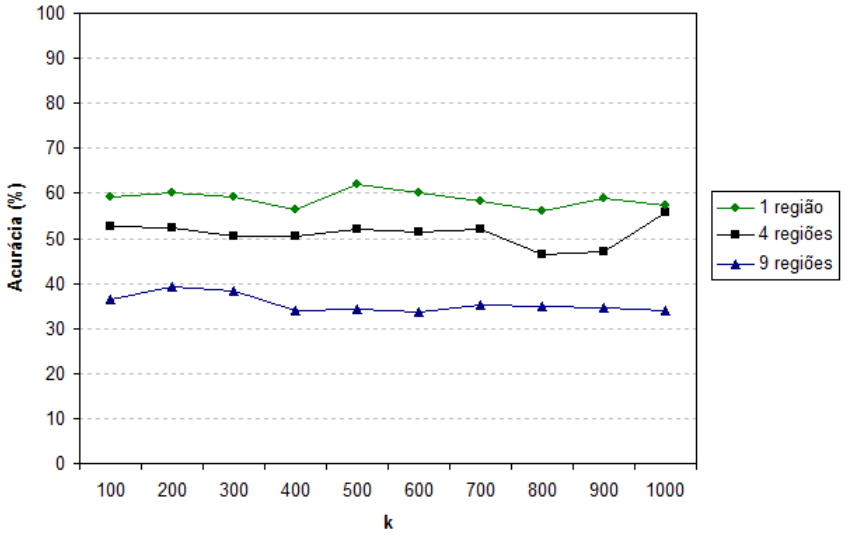

Figura 4.9. Gráfico dos resultados do experimento que constrói o vetor de características da imagem pela concatenação dos histogramas das regiões nas quais a imagem foi dividida. Foram registradas, para cada k, a melhor acurácia obtida.

Jiang et al. [2007]. Os autores argumentam que simplesmente contar os pontos de interesse não é uma escolha ótima, por exemplo, dois pontos associados à mesma palavra visual não são necessariamente igualmente similares àquela palavra, o que significa que suas distâncias ao centróide do grupo relacionado à palavra são diferentes. Ignorar sua similaridade com a palavra visual durante a associação dos pesos faz com que a contribuição dos dois pontos de interesse seja a mesma, e portanto mais difícil de afirmar a importância da palavra visual na imagem.

A solução proposta por Jiang et al. [2007] para esse problema foi o método de

soft-weighting (ponderação suave), que para cada ponto de interesse na imagem, em vez de buscar apenas a palavra visual mais próxima, seleciona as N palavras visuais mais próximas. Em um vocabulário com k palavras visuais, utiliza-se um vetor de k dimensões $T = [t_1, t_2, ..., t_p, ..., t_k]$ em que cada componente t_p representa o peso da palavra visual p na imagem tal que:

$$t_p = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{1}{2^{i-1}} sim(j, p), \tag{4.1}$$

onde M_i representa o número de pontos de interesse em que o i-ésimo vizinho mais próximo é a palavra visual p. A medida $sim(j, p)$ representa a similaridade entre o ponto j e a palavra visual p. Nessa equação, a contribuição de um ponto de interesse é dependente de sua similaridade com a palavra p ponderada pelo fator $\frac{1}{2^{i-1}}$.

Para este experimento foi utilizado $N = 1$ (outros valores de N testados inicialmente como 2, 3 e 4 mostraram pior desempenho) e o mesmo conjunto de imagens do experimento da Subseção 4.3.1, com $k = 100$. Duas outras variações da construção do histograma de palavras visuais por meio de *soft-weighting* foram testadas: utilizando vocabulário com agrupamento separado (Subseção 4.3.3) e utilizando vocabulário construído apenas com imagens de recortes de edifícios. Esses experimentos correspondem aos B, D e E do gráfico da Figura 4.10. Não houve diferença representativa da acurácia destes experimentos em relação ao experimento A. Isso pode ser justificado pelo fato de que como o tamanho do vocabulário é reduzido, pontos similares tendem a ser agrupados em um mesmo grupo. O método de *soft-weighting* foi usado por Jiang et al. [2007] em vocabulários de 500 a 10.000 palavras visuais e, segundo os autores, pontos de interesse similares são agrupados em grupos distintos mais freqüentemente à medida que o tamanho do vocabulário aumenta.

Os experimentos D (construção do histograma de palavras visuais por meio de *soft-weighting* com vocabulário formado por agrupamento separado) e E (construção do histograma de palavras visuais por meio de *soft-weighting* com vocabulário formado apenas com imagens de recortes de edifícios) testaram a possibilidade de uma imagem ser descrita por um vocabulário relacionado apenas a edifícios. Essa idéia foi inspirada pelo trabalho de Perronnin [2008], que utilizou uma abordagem de vocabulários adaptados e universais, a partir da consideração de que uma imagem de uma classe pode ser descrita mais adequadamente pelo vocabulário adaptado dessa classe específica do que pelo vocabulário universal, que descreve o conteúdo de todas as classes. No caso dos experimentos D e E, o desempenho foi semelhante, e isso pode ser explicado pelo

fato de que como a classificação é baseada na detecção de apenas um tipo de objeto, as palavras visuais relacionadas a esse objeto são suficientemente discriminantes, ou seja, as palavras visuais de objetos predominantes em imagens que não sejam de edifícios tendem a ter pouca influência na classificação.

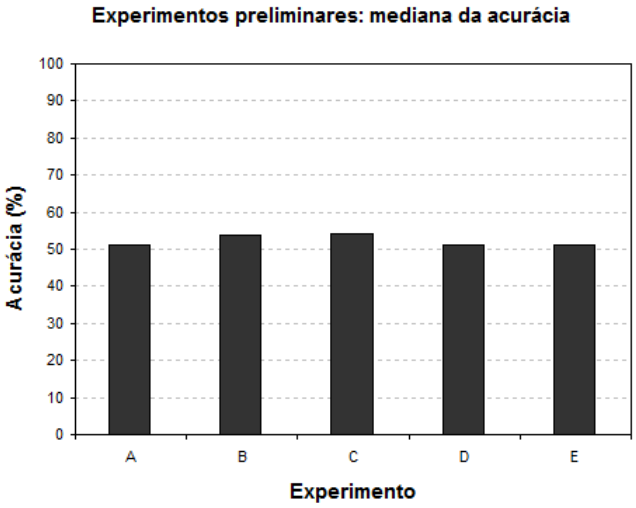

Figura 4.10. Gráfico dos resultados dos experimentos preliminares. O melhor resultado de cada dobra foi obtido, registrando-se a mediana das 5 dobras utilizadas. A = Método sem alterações; B = Construção do histograma de palavras visuais por meio de *soft-weighting*; C = Construção do vocabulário por meio de agrupamento separado; D = B + C; E = Vocabulário formado apenas por recortes de edifícios + B.

4.3.6 Escolha do *kernel*

Embora não seja muito claro qual o *kernel* mais eficaz no contexto da classificação visual, a maioria dos trabalhos nessa área utilizam o *kernel* linear [Nowak et al., 2006; Jurie & Triggs, 2005] ou o Gaussiano [Jiang et al., 2007]. Neste experimento é investigado o impacto da escolha desses dois *kernels* na classificação de imagens do APM.

O mesmo conjunto de imagens do experimento da Subseção 4.3.1 foi utilizado. O tamanho do vocabulário escolhido foi de 100 palavras visuais. Como os melhores parâmetros do *kernel* do classificador SVM para o problema não são conhecidos, é necessário realizar uma busca no espaço de parâmetros com o objetivo de identificar aqueles que possibilitam prever com a melhor acurácia a classe de imagens desconhecidas. Para este experimento, os parâmetros são selecionados por meio da validação cruzada com 3 dobras no conjunto de treino e uma pesquisa em grade (*grid-search*) pelos melhores parâmetros (C e σ para o *kernel* Gaussiano e C para o linear) de cada dobra. Uma varredura no espaço de parâmetros é realizada variando-se os parâmetros

Tabela 4.1. Resultados dos experimento preliminar para definir o *kernel* a ser utilizado. A tabela apresenta a acurácia da classificação, utilizando um intervalo com confiança de 95%.

Kernel	**Acurácia**
Linear	$37,8\% - 39,2\%$
Gaussiano	$43,7\% - 45,5\%$

a partir do limite inferior 0,0002 e até o limite superior 4.000, multiplicando o limite inferior por potências crescentes de 2.

As imagens foram divididas em 5 dobras e para cada *kernel*, os experimentos foram repetidos 30 vezes, (variando-se a composição das dobras aleatoriamente) e calculou-se um intervalo com confiança de 95% da acurácia da classificação. Os resultados são mostrados na Tabela 4.1. Os intervalos calculados demonstram que a melhor escolha para a base experimental é o *kernel* Gaussiano, pois embora próximos, apresentam diferença significativa. De acordo com Jiang et al. [2007], em geral os *kernels* Gaussianos tem melhor desempenho que os lineares. Isso ocorre provavelmente devido ao fato de que palavras visuais estão correlacionadas umas com as outras, e não são linearmente separáveis.

4.4 Experimentos finais

Os experimentos finais foram realizados em dois grupos, utilizando a base experimental com 600 imagens das quais 60 foram retiradas aleatoriamente para compor o conjunto de imagens para vocabulário. Das variações testadas dos experimentos preliminares, conclui-se que a configuração mais adequada para realizar os experimentos finais inclui:

- Um vocabulário com 100 palavras visuais ($k = 100$) obtido pelo agrupamento com *k-means* utilizando-se os descritores SIFT de todos os pontos de interesse extraídos das imagens para vocabulário (totalizando aproximadamente 23.000 descritores);

- Redução dos descritores para 30 dimensões, utilizando-se também todos os descritores das imagens para vocabulário para determinar a matriz de características do PCA;

- O histograma de palavras visuais construído por contagem simples; e

- O classificador SVM com *kernel* Gaussiano.

As 540 imagens que restaram da base experimental foram usadas para realizar validação cruzada com 5 dobras, procedimento que foi repetido 30 vezes, variando-se a composição das dobras aleatoriamente. Para cada execução, a média da acurácia das 5 dobras foi calculada. Com as médias das 30 execuções obteve-se um intervalo com confiança de 95%.

Para os dois grupos de experimentos, os parâmetros são selecionados de maneira semelhante ao experimento preliminar relatado na Subseção 4.3.6.

No primeiro grupo de experimentos o classificador foi treinado com as três classes. Em cada dobra, o conjunto de treino é balanceado retirando-se as imagens de edifícios e não-edifícios a mais que o número de imagens intermediárias. O resultado foi uma taxa média de acertos para a classe edifício de 51,8%, para a classe não-edifício de 58,8% e para a classe intermediária de 37,2% (a acurácia média foi de 49,3%).

Como as imagens rotuladas como intermediárias contêm uma mistura de elementos das duas outras classes, edifícios e não-edifícios, o modelo construído pelo classificador SVM é incapaz de discriminar com sucesso as classes, tendendo a causar confusão na classificação, por exemplo, em imagens que os usuários classificaram como edifícios mas que contém também na cena vegetação.

No segundo grupo de experimentos, o classificador foi treinado apenas com imagens das classes edifício e não-edifício. As imagens pertencentes à classe intermediária foram retiradas do conjunto de treino, que foi balanceado para obter um número semelhante de imagens da classe edifício e não-edifício. Isso foi realizado pela seleção aleatória de um número de imagens de não-edifícios para ser removido das dobras, pois o número de imagens de edifícios é menor. Essa nova configuração experimental resultou em uma taxa média de acertos de 73.1% para a classe edifício e 69.8% para a classe não-edifício. A avaliação foi realizada pela porcentagem de imagens classificadas corretamente das classes edifício e não-edifício individualmente, dentro de um intervalo com confiança de 95%.

Os resultados dos dois grupos de experimentos são resumidos na Tabela 4.2. A confusão entre as imagens é reduzida significativamente quando as imagens da classe intermediária são retiradas do conjunto de treino, ainda que as imagens possam conter objetos de ambas classes, embora em menor proporção.

Esses resultados sugerem que a classe intermediária confunde o classificador. Isso pode ser melhor compreendido pelos exemplos da Figura 4.11, que mostra imagens rotuladas pelos usuários como intermediárias. As imagens dessa classe apresentam dificuldades de serem classificadas tanto pelos usuários quanto pelo classificador, pois

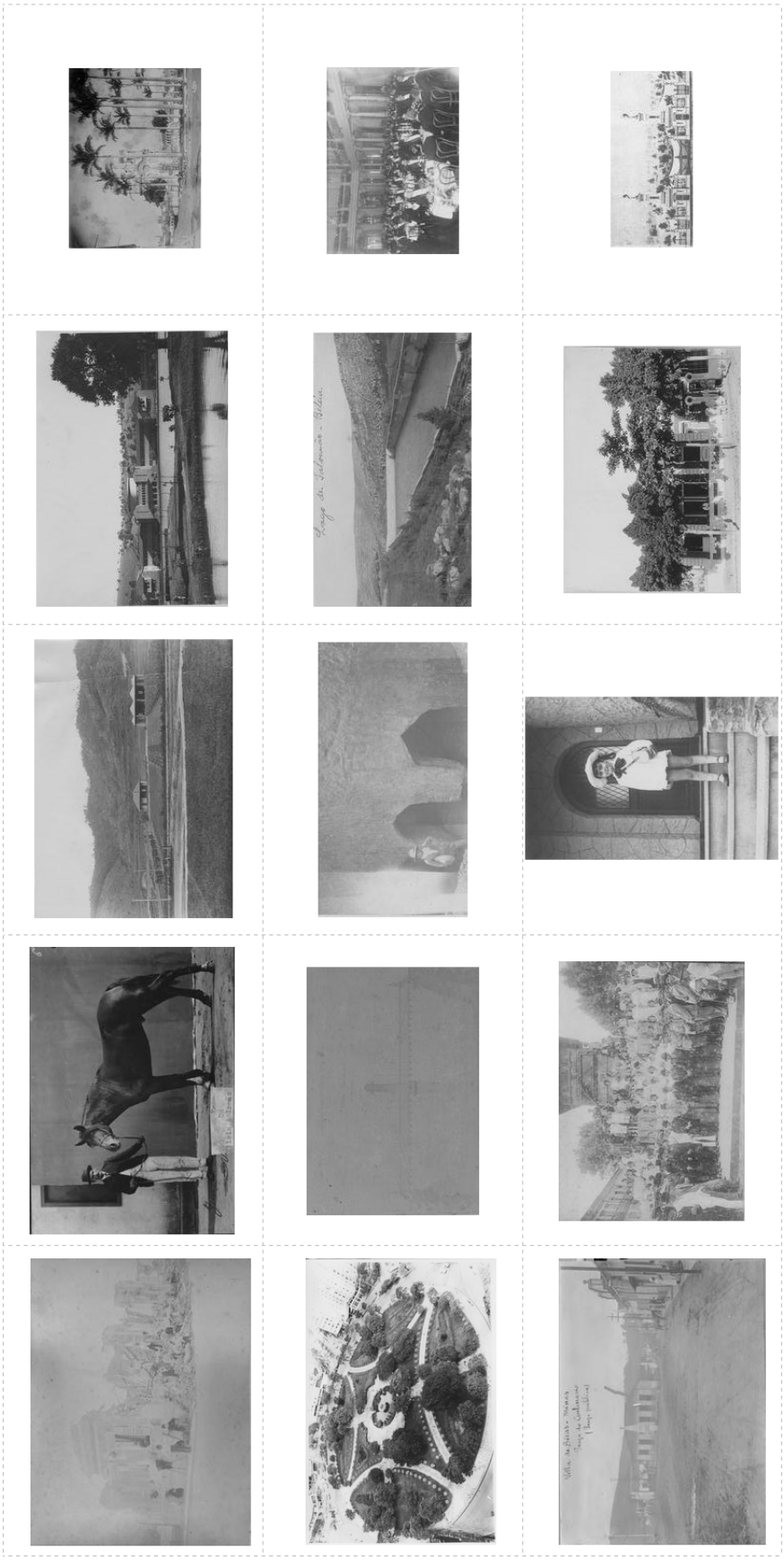

Figura 4.11. Exemplos de imagens da base experimental rotuladas como intermediárias pelos usuários. As imagens da primeira linha são aquelas que três usuários escolheram classes diferentes, sendo necessário um quarto usuário para efetuar o desempate. As imagens da última coluna foram classificadas como edifícios pelo classificador.

Tabela 4.2. Resultados da classificação dos dois grupos de experimentos finais. A tabela mostra a taxa de acerto das classes, utilizando um intervalo com confiança de 95%.

Experimento	Edifício	Não-edifício	Intermediária
1	$50,2\% - 53,4\%$	$57,5\% - 60,1\%$	$34,9\% - 39,6\%$
2	$71,6\% - 74,6\%$	$69,0\% - 70,5\%$	-

Tabela 4.3. Matriz de confusão para uma das rodadas do primeiro grupo de experimentos.

	Edifício	Não-edifício	Intermediária
Edifício	$52,2\%$	$21,7\%$	$26,1\%$
Não-edifício	$17,9\%$	$59,7\%$	$22,4\%$
Intermediária	$23,5\%$	$41,2\%$	$35,3\%$

não pertencem claramente a nenhuma das duas classes (edifícios e não-edifícios). Embora seja possível identificar visualmente construções em algumas imagens da figura, são imagens consideradas pelos usuários como intermediárias, por ocuparem uma porção pouco significativa da imagem, por haver oclusão severa ou por apresentarem pouca nitidez.

A Tabela 4.3 e a Tabela 4.4 mostram as matrizes de confusão de uma das 30 rodadas de cada grupo de experimentos. Os resultados para as imagens intermediárias são apresentados na Tabela 4.4. Em ambos os casos, os números mostram a dificuldade do classificador em discriminar as imagens dessa classe, mesmo no caso mostrado na Tabela 4.3, em que o classificador foi explicitamente treinado levando-se em conta as 3 classes.

O compromisso de se melhorar a taxa de classificação das classes edifício e não-edifício envolve a desvantagem de que as imagens rotuladas como intermediárias serão recuperadas juntamente a essas classes, pois não há nesse caso distinção da classe intermediária. Entretanto, há casos em que alguns exemplos de imagens rotuladas como intermediárias apresentam edifícios na cena, que são detectados pelo classificador (Figura 4.11, exemplos da terceira linha).

Tabela 4.4. Matriz de confusão para uma das rodadas do segundo grupo de experimentos.

	Edifício	**Não-edifício**
Edifício	$73,9\%$	$26,1\%$
Não-edifício	$28,4\%$	$71,6\%$

Tabela 4.5. Porcentagem de imagens anotadas como intermediárias que foram classificadas como edifícios e não-edifícios no segundo grupo de experimentos.

	Edifício	**Não-edifício**
Intermediária	$58,8\%$	$41,2\%$

4.5 Considerações finais

Os experimentos executados neste capítulo mostraram que a utilização de algoritmos em uma base de imagens históricas não é uma tarefa trivial. Das variações testadas dos experimentos preliminares, chegou-se a uma configuração final mais adequada para realizar a detecção de edifícios utilizando a abordagem de histogramas de palavras visuais, que mostrou-se robusta a essas variações por não produzir resultados significativamente distintos entre elas. Para avaliar o desempenho da abordagem, um *ground-truth* foi construído de acordo com a metodologia proposta.

Como mencionado anteriormente (Seção 2.4), não foram encontradas abordagens semelhantes na literatura que lidassem com imagens de fotografias históricas. As abordagens apresentadas por Jurie & Triggs [2005] e Iqbal & Aggarwal [2002b] foram as mais similares à deste trabalho embora os autores trabalharam com fotografias recentes que apresentam boa qualidade visual. As taxas de acerto reportadas por esses trabalhos para a classe edifício foram de 92,0% e 77,6%, respectivamente. Uma comparação grosseira desses valores com aqueles atingidos neste trabalho indicam o potencial da abordagem de histograma de pontos de interesse proposta, mesmo sob ruído elevado e baixa qualidade das imagens.

Capítulo 5

Conclusão

Este trabalho apresentou a aplicação de uma abordagem de histograma de palavras visuais em imagens de fotografias históricas com o objetivo de detectar edifícios nessas imagens. As imagens, que são provenientes da base de fotografias digitalizadas do APM, apresentam variados níveis de degradação, causada pela ação do tempo ou danos sofridos. Essas características desafiam a utilização de algoritmos de Visão Computacional e motivaram a execução de diversos experimentos para avaliar o desempenho da abordagem nessas condições. Realizou-se também a aplicação de uma metodologia proposta para obter um *ground-truth* das imagens para avaliar o algoritmo.

Como a representação de histogramas de palavras visuais tem sido utilizada na literatura em vários cenários de detecção e classificação e provado ser robusta nessas tarefas, optou-se por utilizá-la no problema de detecção de edifícios, com uma implementação baseada em descritores SIFT. Observou-se nos experimentos realizados que o método apresenta níveis de robustez tais, que as variações sugeridas não foram capazes de produzir resultados significativamente distintos.

Os resultados experimentais indicam o potencial do método proposto para a tarefa de detecção de edifícios, mesmo com a baixa qualidade visual das imagens da base utilizada. Uma das contribuições deste trabalho é a aplicação da abordagem de histogramas de palavras visuais em uma base com características difíceis, pois até o limite do nosso conhecimento, todos os trabalhos encontrados na literatura utilizam bases de imagens de fotografias atuais. Os principais resultados deste trabalho foram publicados em uma conferência nacional [Batista et al., 2009b] e em uma conferência latino-americana [Batista et al., 2009a].

O desenvolvimento de técnicas que melhorem a indexação e o acesso ao acervo do APM são importantes para valorização do acervo digital da instituição, que abriga materiais históricos de grande valor para a sociedade. Alguns exemplos de trabalhos já

realizados nesse acervo são a indexação de pessoas baseada na detecção e identificação facial [Lopes et al., 2008a], a identificação automática dos tipos de impressões fotográficas [Oliveira et al., 2006] e a recuperação por imagem-exemplo robusta a distorções [Valle et al., 2006]. Nesse contexto, este trabalho irá cooperar com o desenvolvimento da recuperação de imagens que contém edifícios, informação que também pode ser útil para separar fotografias de áreas rurais ou urbanas.

Como trabalhos futuros, pretende-se amostrar mais pontos de interesse nas imagens. De acordo com Jiang et al. [2007] e Nowak et al. [2006], um dos principais fatores limitantes do desempenho de métodos baseados em pontos de interesse é que eles não são capazes de amostrar densamente o suficiente para produzir os melhores resultados. Um maior número de descritores pode significar mais informação discriminativa para a classificação, o que pode ser alcançado por uma amostragem densa, como utilizado no trabalho de Dalal & Triggs [2005]. Outra alteração na abordagem seria a utilização de um método mais eficaz para agrupamento de dados com alta dimensionalidade para substituir o *k-means* na etapa de construção do vocabulário visual.

Como mencionado no Capítulo 1, uma das principais vantagens da abordagem de histogramas de palavras visuais é ser genérica pois é capaz de lidar com vários tipos de objetos simultaneamente e pode ser facilmente estendida para novos tipos de objetos, o que motiva a utilização da abordagem ou variações dela para detecção de outros objetos na base do APM. Nesse mesmo capítulo foram listados alguns objetos que seriam interessantes para identificação nas imagens do acervo, tais como placas e linhas férreas, que poderiam ser analisados em um trabalho futuro.

Referências Bibliográficas

Agarwal, S.; Awan, A. & Roth, D. (2004). Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1475–1490.

Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, 6:355–385.

Batista, N. C.; Lopes, A. & Albuquerque Araujo, A. (2009a). Detecção de edifícios em fotografias históricas utilizando vocabulários visuais. In *Anais da Conferência Latino Americana de Informática*. (Aceito para publicação).

Batista, N. C.; Lopes, A. & Albuquerque Araujo, A. (2009b). Detecting buildings in historical photographs using bag-of-keypoints. In *Proceedings of the Brazilian Symposium on Computer Graphics and Image Processing*. (Aceito para publicação).

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Chang, C.-C. & Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Disponível em `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. Acessado em 01 de junho de 2009.

Chang, S.-F.; Ma, W.-Y. & Smeulders, A. (2007). Recent advances and challenges of semantic image/video search. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pp. 1205–1208.

Chen, C.; Wactlar, H. D.; Wang, J. Z. & Kiernan, K. (2005). Digital imagery for significant cultural and historical materials. *International Journal on Digital Libraries*, 5(4):275–286.

Cord, M. & Declercq, D. (2001). Three-dimensional building detection and modeling using a statistical approach. *IEEE Transactions on Image Processing*, 10(5):715–723.

Cruz, B. F. (2008). Aplicação da técnica sift na identificação de olhos humanos. Dissertação de mestrado, Universidade Estadual do Rio de Janeiro, Nova Friburgo.

Csurka, G.; Dance, C.; Willamowski, J.; Fan, L. & Bray, C. (2004). Visual categorization with bags of keypoints. In *Proceedings of the Workshop on Statistical Learning in Computer Vision*, pp. 1–22.

da Silva, R. R. G. (2002). *Digitalização de acervos fotográficos públicos e seus reflexos institucionais e sociais: tecnologia e consciência no universo digital.* Tese de doutorado, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pp. 886–893, Los Alamitos, CA, USA. IEEE Computer Society.

Del Bimbo, A. (1999). *Visual information retrieval.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Ding, C.; He, X.; Zha, H. & Simon, H. (2002). Adaptive dimension reduction for clustering high dimensional data. In *Proceedings of the IEEE International Conference on Data Mining*, pp. 147–154.

Farquhar, J. D.; Szedmak, S.; Meng, H. & Shawe-Taylor, J. (2005). Improving "bag-of-keypoints" image categorisation: Generative models and pdf-kernels. Technical report, Image Speech and Intelligent Systems, Department of Electronics and Computer Science, University of Southampton, Southampton, UK.

Fergus, R.; Fei-Fei, L.; Perona, P. & Zisserman, A. (2005). Learning object categories from google's image search. *Proceedings of the IEEE International Conference on Computer Vision*, 2:1816–1823.

Fish, D. A.; Brinicombe, A. M.; Pike, E. R. & Walker, J. G. (1995). Blind deconvolution by means of the richardson-lucy algorithm. *Journal of the Optical Society of America A*, 12(1):58–65.

Forsyth, D. A. & Ponce, J. (2006). *Computer Vision: A Modern Approach.* Prentice Hall of India.

Fradkin, D. & Muchnik, I. (2005). Support vector machines for classification. In Abello, J. & Cormodeg, G., editores, *Discrete Methods in Epidemiology*, volume 70

of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pp. 13–20.

Freund, Y. & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *Proceedings of the International Conference on Machine Learning*, pp. 148–156. Morgan Kaufmann.

Gevers, T. & Smeulders, A. W. M. (2004). Content-based image retrieval: An overview. In Medioni, G. & Kang, S. B., editores, *Emerging Topics in Computer Vision*, IMSC Press Multimedia Series, chapter 8. Prentice Hall, 1 edição.

Gonzalez, R. C. & Woods, R. E. (2003). *Processamento de Imagens Digitais*. Editora Edgard Blücher, primeira edição.

Hanbury, A. (2008). A survey of methods for image annotation. *Journal of Visual Languages & Computing*, 19(5):617–627.

Hsu, C.-W.; Chang, C.-C. & Lin, C.-J. (2009). *A Practical Guide to Support Vector Classification*. Taipei, Taiwan. Disponível em `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`. Acessado em 22 de junho de 2009.

Hu, Y.; Jang, D. S.; Park, J. H.; Cho, S. I. & Lee, C. W. (2008). Real-time apartment building detection and tracking with adaboost procedure and motion-adjusted tracker. *ETRI JOURNAL*, 30(2):338–340.

Iqbal, Q. & Aggarwal, J. (2002a). Cires: A system for content-based retrieval in digital image libraries. In *Proceedings of 7th International Conference on Control, Automation, Robotics and Vision*, volume 1, pp. 205–210.

Iqbal, Q. & Aggarwal, J. (2002b). Retrieval by classification of images containing large manmade objects using perceptual grouping. *Pattern Recognition Journal*, 35(7):1463–1479.

Iqbal, Q. & Aggarwal, J. K. (1999). Using structure in content-based image retrieval. In *Proceedings of the Signal and Image Processing Conference*, pp. 129–133.

Iqbal, Q. & Aggarwal, J. K. (2001). Perceptual grouping for image retrieval and classification. In *Proceedings of the IEEE Computer Society Workshop on Perceptual Organization in Computer Vision*.

Iqbal, Q. & Aggarwall, J. (1999). Applying perceptual grouping to content-based image retrieval: building images. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:42–48.

Jain, A. K. (2009). Data clustering : 50 years beyond k-means. Technical Report TR-CSE-09-11, Michigan State University, East Lansing, Michigan.

Jain, A. K.; Murty, M. N. & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis: techniques for experimental design, measurement, simulation, and modeling*. Wiley Computer Publishing.

Jiang, Y.-G.; Ngo, C.-W. & Yang, J. (2007). Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 494–501, New York, NY, USA. ACM.

Jung, C. R. & Schramm, R. (2004). Rectangle detection based on a windowed hough transform. In *Proceedings of the Symposium on Computer Graphics and Image Processing*, pp. 113–120. IEEE Computer Society.

Jurie, F. & Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, pp. 604–610.

Karantzalos, K. & Paragios, N. (2009). Recognition-driven two-dimensional competing priors toward automatic and accurate building detection. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1):133–144.

Kennedy, L. S.; Chang, S.-F. & Kozintsev, I. V. (2006). To search or to label?: predicting the performance of search-based automatic image classifiers. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, pp. 249–258, New York, NY, USA. ACM.

Kim, Y.; Lee, K.; Choi, K. & Cho, S. I. (2006). Building recognition for augmented reality based navigation system. In *Proceedings of the IEEE International Conference on Computer and Information Technology*, pp. 131–131.

Kumar, S. & Hebert, M. (2003). Man-made structure detection in natural images using a causal multiscale random field. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pp. 119–126.

Kundur, D. & Hatzinakos, D. (1996). Blind image deconvolution. *IEEE Signal Processing Magazine*, 13(3):43–64.

Lazebnik, S.; Schmid, C. & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178, Washington, DC, USA. IEEE Computer Society.

Li, Y.; Shapiro, L. & Bilmes, J. (2005). A generative/discriminative learning algorithm for image classification. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pp. 1605–1612.

Li, Y. & Shapiro, L. G. (2002). Consistent line clusters for building recognition in cbir. In *Proceedings of the International Conference on Pattern Recognition*, volume 3, pp. 952–956, Washington, DC, USA. IEEE Computer Society.

Lopes, A.; Oliveira, C. & Albuquerque Araujo, A. (2008a). Face recognition aiding historical photographs indexing using a two-stage training scheme and an enhanced distance measure. In *Proceedings of the Brazilian Symposium on Computer Graphics and Image Processing*, pp. 11–18.

Lopes, A. P. B.; Flam, D. L.; Batista, N. C.; de Avila, S. E. F.; Almeida, J. M. & de A. Araújo, A. (2008b). Automatic frame extraction for improving content-based image retrieval of historical photographs. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pp. 44–50, Vila Velha - ES, Brasil. ACM.

Lorena, A. C. & de Carvalho, A. C. P. L. F. (2003). Introdução às máquinas de vetores suporte (support vector machines). Technical Report 192, Universidade de São Paulo, São Carlos.

Lorena, A. C. & de Carvalho, A. C. P. L. F. (2007). Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, 14(2):43–67.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pp. 1150–1157.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Marszalek, M. & Schmid, C. (2006). Spatial weighting for bag-of-features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:2118–2125.

Matlab (2009). Image deblurring. In *Matlab Image Processing Toolbox User's Guide*, chapter 13.

Mayer, H. (1999). Automatic object extraction from aerial imagery—a survey focusing on buildings. *Computer Vision and Image Understanding*, 74(2):138–149.

Noronha, S. & Nevatia, R. (2001). Detection and modeling of buildings from multiple aerial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(5):501–518.

Nowak, E.; Jurie, F. & Triggs, B. (2006). Sampling strategies for bag-of-features image classification. In *Proceedings of the European Conference on Computer Vision*, pp. 490–503. Springer.

Oliveira, C. J. S.; Batista, N. C.; de Albuquerque Araújo, A. & de Souza Pio, J. L. (2006). Photographic prints identification. In Enyedi, B. & Reichardt, A., editores, *Proceddings of the International Conference on Systems, Signals and Image Processings and Semantic Multimodal Analysis of Digital Media*, pp. 231–234, Budapest, Hungary. Harlequin, s.r.o. Kosice.

Oliveira, C. J. S. & de Albuquerque Araújo, A. (2004). Levantamento dos projetos que envolvem bases de imagens digitais do arquivo público mineiro. Technical Report RT.DCC.001/2004, Núcleo de Processamento Digital de Imagens da UFMG (NPDI/DCC/ICEx/UFMG), Belo Horizonte, Brasil.

Pedrini, H. & Schwartz, W. R. (2008). *Análise de Imagens Digitais: Princípios, Algoritmos e Aplicações*. Thomson Learning.

Perronnin, F. (2008). Universal and adapted vocabularies for generic visual categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1243–1256.

Portugal, M. S. (2005). Notas introdutórias sobre o princípio de máxima verossimilhança: Estimação e teste de hipóteses. Technical Report TD 04, Departamento de Ciências Econômicas, Universidade Federal do Rio Grande do Sul, Porto Alegre.

Saeedi, P. & Zwick, H. (2008). Automatic building detection in aerial and satellite images. In *Proceedings of the International Conference on Control, Automation, Robotics and Vision*, pp. 623–629.

Santos, R. J. (2007). *Um Curso de Geometria Analítica e Álgebra Linear*. Imprensa UFMG.

Shao, H. & Gool, L. V. (2003). Zubud-zurich buildings database for image based recognition. Technical Report 260, ETH Zürich/Swiss Federal Institute of Technology, Zürich.

Sharma, G.; Chaudhury, S. & Srivastava, J. (2008). Bag-of-features kernel eigen spaces for classification. In *Proceedings of the International Conference on Pattern Recognition*, pp. 1–4.

Simões, A. S. & Costa, A. H. R. (2002). Utilizando processos gaussianos para a segmentação de imagens monocromáticas. In *Anais do Seminário de Computação da Universidade Regional de Blumenau*, pp. 177–188, Blumenau, SC.

Smith, L. I. (2002). A tutorial on principal components analysis. Disponível em `http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf`. Acessado em 01 de junho de 2009.

Torres, R. d. S. & Falcão, A. X. (2006). Content-based image retrieval: Theory and applications. *Revista de Informática Teórica e Aplicada*, 13(2):161–185.

Trinh, H.-H. & Jo, K.-H. (2006a). Image-based structural analysis of building using line segments and their geometrical vanishing points. In *Proceedings of the SICE-ICASE International Joint Conference*, pp. 566–571.

Trinh, H.-H. & Jo, K.-H. (2006b). Line segment-based facial appearance analysis for building image. In *Proceedings of the International Forum on Strategic Technology*, pp. 332–335.

Trinh, H.-H.; Kim, D.-N. & Jo, K.-H. (2007). Urban building detection by visual and geometrical features. In *Proceedings of the International Conference on Control, Automation and Systems*, pp. 1779–1784.

Trucco, E. & Verri, A. (1998). *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, Upper Saddle River, NJ, USA.

Valle, E. (2003). Sistemas de informação multimídia na preservação de acervos permanentes. Dissertação de mestrado, Universidade Federal de Minas Gerais, Belo Horizonte.

Valle, E.; Cord, M. & Philipp-Foliguet, S. (2006). Content-based retrieval of images for cultural institutions using local descriptors. In *Geometric Modeling and Imaging–New Trends*, pp. 177–182.

Wang, J. Z.; Grieb, K.; Zhang, Y.; chih Chen, C.; Chen, Y. & Li, J. (2006). Machine annotation and retrieval for digital imagery of historical materials. *International journal on digital libraries*, 6(1):18–29.

Yuan, X. & Li, C.-T. (2006). Cbir approach to building image retrieval based on linear edge distribution. In *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*, p. 95, Los Alamitos, CA, USA. IEEE Computer Society.

Zhou, H. & Suter, D. (2008). Improved building detection by gaussian processes classification via feature space rescale and spectral kernel selection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–6.

Apêndice A

Glossário de termos

1. Aprendizado de Máquina: as técnicas de Aprendizado de Máquina empregam um princípio de inferência denominado indução, no qual obtém-se conclusões genéricas a partir de um conjunto particular de exemplos. O aprendizado indutivo pode ser dividido em dois tipos principais: supervisionado e não-supervisionado [Lorena & de Carvalho, 2007].

2. Autovalor e autovetor: Seja A uma matriz $n \times n$. Conforme definição de Santos [2007], um número real λ é chamado autovalor (real) de A, se existe um vetor não nulo $V = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$ de $\mathbb{R}$, tal que

$$AV = \lambda V.$$

   Um vetor não nulo que satisfaça a equação acima é chamado de autovetor de A.

3. *Boosting*: método utilizado para aprendizagem supervisionada que pode reduzir significativamente o erro de algoritmos de aprendizado denominados fracos, cujo desempenho é pobre. Consiste em combinar algoritmos de aprendizado fracos para produzir um classificador composto, denominado forte, que apresente melhor desempenho do que qualquer dos classificadores fracos [Freund & Schapire, 1996].

4. Características da imagem: em Visão Computacional, o termo característica refere-se a duas possíveis entidades: (i) uma propriedade global de uma imagem ou parte dela, por exemplo o nível de cinza médio; ou (ii) uma parte da imagem com alguma propriedade especial, por exemplo um círculo, uma linha ou uma região texturizada em uma imagem de intensidade [Trucco & Verri, 1998].

5. Características de Haar: características cujo valor é calculado de forma semelhante aos coeficientes de uma transformada de Haar e extraídas de uma imagem por meio de máscaras, tais como a da Figura A.1 [Hu et al., 2008].

Figura A.1. Máscaras para extração de características de Haar. Figura retirada de Hu et al. [2008].

6. Classificador dos k vizinhos mais próximos (*k-nearest neighbor*): este método baseia-se na heurística que considera que pontos de exemplo próximos a um ponto cuja classe não é conhecida podem indicar a classe deste ponto. Dessa forma, um ponto é classificado usando a classificação dos k exemplos mais próximos, cujas classes são conhecidas, sendo a classe com o maior número de votos a escolhida [Forsyth & Ponce, 2006].

7. Descritor de imagem: no contexto de RIBC, a construção de um descritor de uma imagem é caracterizada por: (i) um algoritmo para extração das características de um imagem (por exemplo cor, textura, forma, etc), codificando-as em vetores de características; e (ii) uma medida de similaridade para comparar duas imagens [Torres & Falcão, 2006].

8. Histograma: contém a distribuição de probabilidades de dados, representado pela freqüência dos dados em determinados intervalos. Como exemplo, tem-se o histograma de cor, que é o método mais tradicional para descrever características de baixo-nível das imagens. É obtido pela discretização das cores da imagem e a contagem de quantos *pixels* pertencem a cada cor [Del Bimbo, 1999]. O histograma de uma imagem corresponde à distribuição dos níveis de cinza da imagem, o qual pode ser representado por um gráfico indicando o número de pixels na imagem para cada nível de cinza (*bin*) [Pedrini & Schwartz, 2008].

9. Imagem digital: pode ser considerada como sendo uma matriz cujos índices de linhas e colunas identificam um ponto na imagem, e o correspondente valor do elemento da matriz identifica o nível de cinza naquele ponto [Gonzalez & Woods, 2003].

10. Máxima Verossimilhança: método de inferência estatística que fornece um procedimento de obtenção de estimativas de parâmetros desconhecidos com base em conjuntos de dados experimentais independentes [Portugal, 2005].

11. Recuperação de informação: é o estudo de sistemas que recuperam itens de coleções usando vários tipos de informações [Forsyth & Ponce, 2006].

12. *Pixel*: cada elemento da matriz de uma imagem digital. São também chamados elementos dda imagem, elementos da figura ou *pels*. *Pixel* ou *pel* são abreviações do inglês *picture elements* [Gonzalez & Woods, 2003].

13. Processos Gaussianos: seja um vetor de k dimensões $x = x_1, x_2, \ldots, x_k$. A distribuição de probabilidade de uma função $y = f(x)$ é definida como um processo gaussiano se, para qualquer seleção de pontos em x, a densidade marginal $P(f(x_1), f(x_2), \ldots, f(x_k))$ for da forma gaussiana. Processos gaussianos podem ser utilizados para aprendizagem por meio de regressão e predição, em que buscase realizar a regressão de determinada função de classificação e sua predição para os demais valores do domínio [Simões & Costa, 2002].

14. Rastreamento (*tracking*): é o problema de gerar uma inferência sobre o movimento de um objeto dada uma seqüência de imagens [Forsyth & Ponce, 2006]. Problemas típicos de rastreamento utilizam um modelo do movimento do objeto e um determinado conjunto de medidas de uma seqüência de imagens, tais como a posição de alguns pontos ou os momentos de algumas regiões da imagem.

15. Realidade aumentada: é um ambiente que envolve tanto realidade virtual como elementos do mundo real. Neste paradigma, o usuário está imerso em um ambiente sintético e pode ver o mundo real com objetos virtuais superimpostos ou colocados em composição com o mundo real [Azuma, 1997].

16. Transformada de Hough: é um método utilizado para detectar estruturas lineares em imagens. A aplicação da transformada de Hough em um conjunto de pontos (x_i, y_i) de uma borda resulta em uma função bidimensional $C(\rho, \theta)$ que representa o número de pontos de borda que satisfazem a equação linear $\rho = x \cos\theta + y \sin\theta$, em que ρ é distância normal e θ é o ângulo normal de uma linha reta. Máximos locais de C podem ser utilizados para detectar segmentos de linha que passam por pontos de borda [Jung & Schramm, 2004].

17. Textura: descritor de uma imagem ou de uma região que fornece medidas de propriedades tais como suavidade, rugosidade e regularidade [Gonzalez & Woods, 2003]. Imagens de textura geralmente consistem em padrões organizados de subelementos regulares, chamados *textons* [Forsyth & Ponce, 2006].