Paulo Sérgio Silva Rodrigues
Orientador: Arnaldo de Albuquerque Araújo

Um Modelo Bayesiano Combinando Análise Semântica Latente e Atributos Espaciais para Recuperação de Informação Visual

Tese apresentada ao Curso de Pós-graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

Belo Horizonte
12 de Março de 2003

Deus meu, faça-me à tua vontade cada vez mais humilde, fraterno e justo.

Agradecimentos

Primeiramente, dedico este trabalho a Deus, que sempre tem me mostrado que, por mais conhecimento que eu adquira, é sempre ele quem me conduz. Quando a situação fugiu de meu controle, foi para mostrar-me que é ele, e mais ninguém, quem guia as nossas vidas.

Dedico também aos Papas João XXIII, Paulo VI, João Paulo I e João Paulo II. O exemplo de vida, de humildade, bondade, amor e dedicação pelo trabalho que lhes designou Deus, me influencia e continua sendo como uma força para cada passo que eu dou.

Entre as pessoas que contribuíram diretamente para a conclusão de minha tese está inicialmente o meu orientador Arnaldo Albuquerque, que sozinho, teve paciência e dedicação comigo, desde o mestrado até o doutorado.

Agradeço imensamente aos meus professores da pós-graduação: Nívio Ziviani, Frederico Campos, Newton Vieira, Claudionor Coelho, Antônio Otávio, Antônio Loureiro e Mário Fernando, que contribuíram passando-me seus conhecimentos. No caso particular do professor Mário, não posso deixar de esquecer as longas horas que dedicou a mim com seus conselhos, os quais sem dúvida influenciaram meus caminhos tanto na tese quanto na vida pessoal.

A seguir, agradeço a diversas pessoas que contribuíram direta ou indiretamente para que eu chegasse ao final de meu doutorado. Os nomes não aparecem em nenhuma ordem específica, fui citando a medida que me vinham à mente:

Silvio Jamil, meu amigo pessoal e colega de área de pesquisa. Silvio esteve sempre do meu lado, dando o seu apoio moral e muitas vezes técnico.

Camillo Oliveira, que ficava atento sempre que possível a novas publicações na área, com o intuito de manter meu trabalho sempre atual.

Leonardo Claudino: pela amizade e discussão técnica que tivemos, que ajudou, sem dúvida, a melhorar meus conhecimentos.

Linnyer Beatriz: pelas palavras sempre oportunas e de conforto que me dedicou.

José Pio: pela força que me deu, pela boa amizade e reconhecimento.

Raquel Mini: pela eterna amizade e por estar ao meu lado em todos os momentos bons e difíceis. À Raquel, eu não só agradeço como também peço perdão por nem sempre compreender o carinho e amizade que sempre dedicou a mim, que muitas vezes obrigou-lhe a dizer palavras duras, mas oportunas, que só contribuíram para o meu crescimento profissional.

Elaine Pimentel: pelos bons conselhos, força, e bons goles que tomamos juntos.

Lucila Ishitani: por acreditar em mim em momentos difíceis e depositar sua confiança

quando me ofereceu uma turma de ciência da computação para lecionar.

Hermes Júnior: pelas longas conversas amistosas que tivemos, aconselhando-nos mutuamente, pela sua amizade, respeito e paciência para comigo, e pelas boas partidas de tênis que jogamos. Hermes, o "Mulão", me ajudou também a descontrair e manter um preparo físico, muitas vezes necessário.

Rodrigo Valle: pela sua imensa amizade e frases irreverentes que, sem dúvida, me ajudaram a descontrair em momentos de tensão.

Vânia Evangelista: pela sua sincera amizade.

Daniel de Moura: pela sua amizade, palavras de força e momentos de alegria. Daniel, que conheço a menos de um ano, é um desses amigos que, desde quando fomos apresentados, parece ser um amigo de infância.

Lena Veiga: pela amizade sincera e bons momentos de descontração. Lena, como Daniel, é minha amiga recente que parece ser desde a infância.

Maria Di Lourdes: pela amizade e imprescindível colaboração tanto na correção do texto quanto no conteúdo técnico. É dela uma parte do texto com relação às definições de redes bayesianas.

Pável Calado: pela amizade, pela ajuda com relação à teoria das redes bayesianas (sem às quais grande parte da tese não seria possível), pelas boas conversas descontraídas e pelo preparo da boa cozinha portuguesa que nos oferecia sempre que se dispunha, com sua paciência e inteligência, a nos oferecer.

Gurvan Huiban: pela amizade, ajuda com o computador muitas vezes e pelas boas cervejas que tomamos juntos, sobre tudo em casa, ao chegarmos.

Silvana Bocanegra: pela sincera amizade e apoio em momentos difíceis. Silvana esteve comigo quando mais precisei, devido a uma saúde um tanto duvidosa.

Daniella Mota: também pela amizade e apoio em momentos difíceis. Dela, eu também não esquecerei a força durante minha convalescência.

Belkiz: pela amizade e ajuda com os periódicos da biblioteca, sem os quais, seria impossível concluir meu trabalho. Sua competência a frente da biblioteca setorial do DCC/ICEx sem dúvida tem contribuído, não só a mim, mas a todos os outros pós-graduandos.

Helvécio: por, como Bel, trabalhar com competência e alegria. A Helvécio eu me dou o direito de agradecer por todos os alunos de pós-graduação.

Antônia: o cafezinho e o chá que faz com tanta competência, sem dúvida contribuíram para amenizar os longos períodos de tempo que passamos concentrados.

D. Oneida Santos: sem a sua estimada colaboração e os cuidados com meus objetos pessoais seria mais difícil, com certeza, minha estada em Belo Horizonte.

Renata: sem Renata, a pós-graduação no DCC-UFMG não teria o porte e reconhecimento que tem. A ela eu também agradeço a paciência não só para comigo, mas para todos os alunos que ela conheceu.

Emília: pela amizade, dedicação e competência no trabalho que faz, e por suportar um monte de "alunos chatos" como eu.

Lúcio França (*in memorian*): Lúcio foi meu amigo pessoal. conversamos longas horas e fizemos planos de trabalhar e ajuda mútua. Infelizmente o Senhor o chamou muito cedo.

Finalmente: agradeço a toda a minha família, que se privou de minha presença e suportou a saudade. À minha mãe e meu pai por me amarem. Não existem palavras que sejam suficientes para expressar todo o carinho e importância que foram fundamentais para tudo o que ocorreu de bom em minha vida.

Resumo

Este trabalho apresenta a implementação e análise de três novas metodologias distintas para extração de características e recuperação de informações visuais em imagens digitais. A primeira é um algoritmo, chamado de GRAS (*Graph Region Arrow Shot*), que utiliza informações do relacionamento espacial entre as regiões da imagem. Essas características são acrescentadas àquelas de métodos clássicos, como histograma de cores, mapa de bordas e informações de textura. A segunda usa decomposição em valor singulares para extrair informações de co-ocorrência de características, transformando o espaco original em um novo espaço de trabalho, o "espaço latente". Finalmente, a terceira metodologia usa um modelo bayesiano para combinar o resultado da ordenação de métodos clássicos, obtendo um melhor desempenho. Para avaliação dos métodos, duas bases de dados são utilizadas: uma natural e outra artificial. Na base natural, que consta de diversas classes de imagens reais, classificadas manualmente, os resultados das novas técnicas foram melhores. Nos experimentos, quando o algoritmo GRAS é usado, obteve-se cerca de 60% de melhora na precisão média.

Abstract

This thesis presents an implementation and analysis of three new methodologies for feature extraction and retrieval of visual information of digital images. The first implemented methodology is a new algorithm, called GRAS (*Graph Region Arrow Shot*), which uses information from the spacial relationship between the image's regions. The GRAS information is added to features of classical methods such as color histograms, edge maps and texture information. The second one uses singular value decomposition to extract information from the image context, transforming the original space in a new space, which is called *latent space*. The third methodology uses a bayesian network model which combines the results from the classical methods, achieving better performance than each method alone. For performance evaluation, two image databases are used: one natural and another artificial. In the natural database, which is composed of several manually classified natural images, the results of the new techniques are better. In the experiments, when the GRAS algorithm is used, the improvements reach up to 60% in terms of average precision.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

As últimas décadas testemunharam um grande desenvolvimento de áreas populares como multimídia, jogos e a *World Wide Web*, e áreas científicas como as que manipulam grandes volumes de dados, compostos cada vez mais de imagens. Este desenvolvimento tem levado à crescente necessidade por soluções para tratar tarefas como visualizar, armazenar, manter, recuperar e transmitir informações de imagens digitais. Comparado com o estado da arte de algumas décadas atrás, é notável o avanço das técnicas para manipular esse tipo de dado. No entanto, ainda há muito a ser feito.

Entre as tarefas de maior demanda na área está a de recuperar informações em imagens com base em seu conteúdo visual, que normalmente é chamada de RIBC (Recuperação de Imagens Baseada no Conteúdo)[1] [Niblack, 1993, Smith e Chang, 1996a]. Essa tarefa é ainda dividida em diversas outras, como no trabalho de [Marsicoli et al., 1997], e podem ser: pré-processamento, segmentação, extração de características, indexação e recuperação das imagens propriamente dita. No entanto, essa divisão pode variar e cada uma em si pode representar um problema que exija muito tempo de pesquisa e esforço computacional. Apesar de ser uma área ramificada, muitos trabalhos clássicos, como os de [Smith e Chang, 1996a],

[1] Esta sigla estende RI, originalmente usada para Recuperação de Informação textual

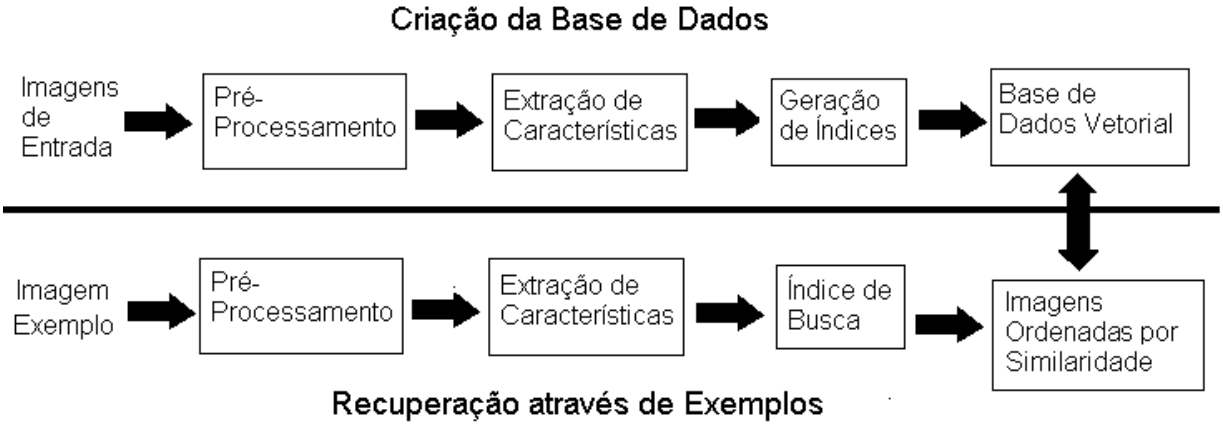


Figura 1.1: Esquema geral de um sistema de recuperação de imagens com base no conteúdo (RIBC) usando uma imagem-exemplo como argumento de busca.

[Niblack, 1993] e [Gudivada e Raghavan, 1995], abordam o problema como um todo, cujos experimentos mostram melhorias em apenas algumas etapas ou em cada uma delas individualmente. Na verdade, a área de RIBC é dividida em duas fases principais, uma *off-line*, onde são gerados índices para cada imagem de entrada, e outra *online*, onde são feitas as buscas propriamente ditas. Por sua vez, tanto a etapa *off-line* quanto a etapa *online* podem ser divididas em diversas etapas menores. Normalmente, a busca é realizada através de uma imagem-exemplo (ou imagem-consulta), que pode ou não já pertencer 'a base de dados. Uma função que minimiza a distância entre o índice gerado para a imagem-exemplo e os índices da base de dados (e conseqüentemente entre a imagem-exemplo e todas as imagens da base de dados) é utilizada como medida de similaridade. Geralmente, o resultado é uma ordenação de todas as imagens da base, da mais similar para a menos similar. A Fig. 1.1 mostra um esquema geral dessa abordagem.

Por esse esquema, a fase *off-line* consta normalmente de processar e indexar toda a base de dados com o intuito de uma melhor eficiência na parte final do processo, na fase *online*. Nesta última etapa, normalmente o usuário pode apresentar como entrada para o sistema um esboço de um desenho que queira recuperar, uma imagem-exemplo do que deseja ou simplesmente uma combinação de ambos. Sistemas em que o usuário apresenta uma imagem-exemplo são chamados de *sistemas baseados em exemplos*. Por

outro lado, sistemas em que o usuário fornece como entrada simplesmente a frase ou palavra relacionados com o tipo de imagem que deseja, não são considerados neste trabalho, pois suas buscas não levam em conta o conteúdo da imagem.

Todas as etapas apresentadas na Fig. 1.1 são bem definidas e podem ser estudadas separadamente. Entretanto, cada uma depende da anterior. Por exemplo, para extrair características, geralmente, é necessário que bons algoritmos de segmentação tenham sido utilizados previamente; a indexação é dependente de quais características estão sendo indexadas; por sua vez, para ser eficiente, a recuperação depende de quais estruturas de dados foram utilizadas para a indexação.

A extração de características, tanto na fase *off-line* quanto na fase *online*, é o processo que tem gerado os maiores desafios na área, e também é dos que mais atenção tem recebido. Os problemas nesse processo vêm desde quais características relevantes extrair até problemas de como extrair e armazená-las. Sabe-se, contudo, que são muitas as características que podem ser indexadas nesse tipo de sistema. No entanto, os trabalhos desenvolvidos até hoje concentram-se em principalmente três — e no máximo em quatro — tipos: cor, forma, textura e relacionamento espacial entre os objetos ou regiões da imagem.

O trabalho apresentado aqui concentra-se apenas na extração de características, tanto incorporando novas às já conhecidas, quanto apresentando uma modelagem de como combiná-las de maneira que sistemas baseados em RIBC obtenham um desempenho melhor com relação à quantidade de imagens recuperadas semelhantes à imagem-exemplo.

1.1.1 O uso da cor, forma e textura em RIBC

Características de cor

Historicamente, tentar recuperar imagens através de seu conteúdo de cor tem sido a estratégia mais usada. Nesse caso, o cálculo do histograma de cores da imagem é, de longe, a técnica que tem gerado o maior volume de trabalhos. Isso é devido principalmente a dois fatores: histogramas são simples de computar e são invariantes às operações lineares da imagem como escala, translação e rotação. Entretanto, histogramas não capturam o

relacionamento espacial entre os objetos ou regiões das imagens: cenas semanticamente diferentes podem possuir histogramas de cores iguais; e cenas semanticamente iguais podem possuir histogramas de cores diferentes. Esse problema é agravado devido ao fato de que, normalmente, uma quantização do espaço de cor da imagem é feita com o objetivo de diminuir o espaço em disco requerido, eliminar ruídos e a quantidade de processamento necessária. Por tudo isso, quantizar pode levar a perdas de informações relevantes. A Fig. 1.2 exibe um exemplo de imagens coloridas (no espaço HSV) e seus histogramas, após uma quantização para 24 valores[2]. Nela, as Figs. 1.2-(a), 1.2-(c) e 1.2-(e), correspondentes às imagens Ia, Ic e Ie, são as imagens originais. Semanticamente falando, Ia e Ic pertencem (ou pelo menos espera-se que assim o sejam) à mesma classe "por-do-sol" e a imagem Ie à classe "carro". Se a idéia aqui for tentar classificar as 3 imagens dessa figura por semelhança, usando uma função de similaridade como, por exemplo, a correlação do ângulo entre seus respectivos histogramas, quanto menor for o ângulo entre eles mais "semelhantes" as imagens serão. Nesse caso particular, os histogramas usados nas Figs. 1.2-(b), 1.2-(d) e 1.2-(f), correspondentes às imagens Ia, Ic e Ie, respectivamente, foram:

- (a) Hb = [392 4230 2099 2037 12 5 8 3 2 2 10 26 14 8 3 2 2 0 116 226 8513 675 3400 4780];
- (b) Hd = [154 941 4308 3324 40 9 36 2 0 2 1 15 1 6 0 4 12 0 32 946 7877 3468 1393 3994];
- (c) Hf = [1637 1802 4164 88 10 1 11 0 0 0 1 7 5 3 3 12 63 1048 10 1487 7358 2514 3635 2706].

Neste caso particular, o ângulo entre os histogramas Hb e Hd — e conseqüentemente a semelhança entre as imagens Ia e Ic — é $\cos^{-1}(0.8811) = 28.22^o$ e entre Hb e Hf — e

[2]Neste trabalho, uma imagem quantizada no espaço HSV contém 18 valores para o *Hue* (matiz, no português, que representa as diferentes cores possíveis), 3 valores para a *Saturation* (Saturação, no português, que representa a quatidade de branco presente na cor) e 3 valores para o *Value* (Valor, no português, que representa a quantidade de luz presente). Essa representação, $18 + 3 + 3 = 24$, chamada de HSV-162, por permitir $18 \times 3 \times 3 = 162$ cores possíveis, foi escolhida por ser a mais próxima do Sistema Visual Humano [Fuertes et al., 2001]

conseqüentemente a semelhança entre as imagens Ia e Ie — é $\cos^{-1}(0.8924) = 26.82^o$. Isso permite dizer que a imagem Ia é semanticamente mais próxima da imagem Ie do que da imagem Ic!

Um outro problema dessa abordagem é o "fator de quantização" (número de cores quantizadas), que influencia diretamente no resultado final da busca. No entanto, o cálculo do fator ideal ainda é um problema em aberto, sendo, portanto, utilizados valores empíricos. Por ser uma técnica simples, bastante conhecida e testada, a abordagem baseada em histograma de cores no espaço HSV foi escolhida neste trabalho como ponto de comparação com os novos modelos que serão apresentados mais tarde.

Características de forma

Estratégias que utilizam o conteúdo de forma da imagem como argumento de indexação podem seguir duas vertentes. A primeira leva em conta a orientação que cada pixel tem em relação a um determinado referencial, criando o que é conhecido na literatura como "mapa de bordas da imagem" [Gonzalez e Woods, 1992]. A segunda identifica o contorno (borda externa) dos objetos ou regiões da imagem, através de algoritmos denominados de "códigos de cadeia" [Gonzalez e Woods, 1992]. A Fig. 1.3 ilustra ambas as estratégias.

O problema do mapa de bordas é semelhante ao problema dos histogramas de cores: imagens semanticamente iguais podem ter "mapas de bordas" distintos, e vice-versa. Entretanto, um dos problemas mais sérios da abordagem baseada em contornos é o fato de ser restrita às imagens que possuem "objetos fechados", como a Fig. 1.3(c). Isso restringe esse tipo de estratégia às imagens simples com objetos bem definidos e fundo homogêneo. Neste trabalho, foi escolhido o mapa de bordas como modelo clássico de representação de características de bordas. Por esse motivo, os métodos novos estudados mais tarde serão comparados ele.

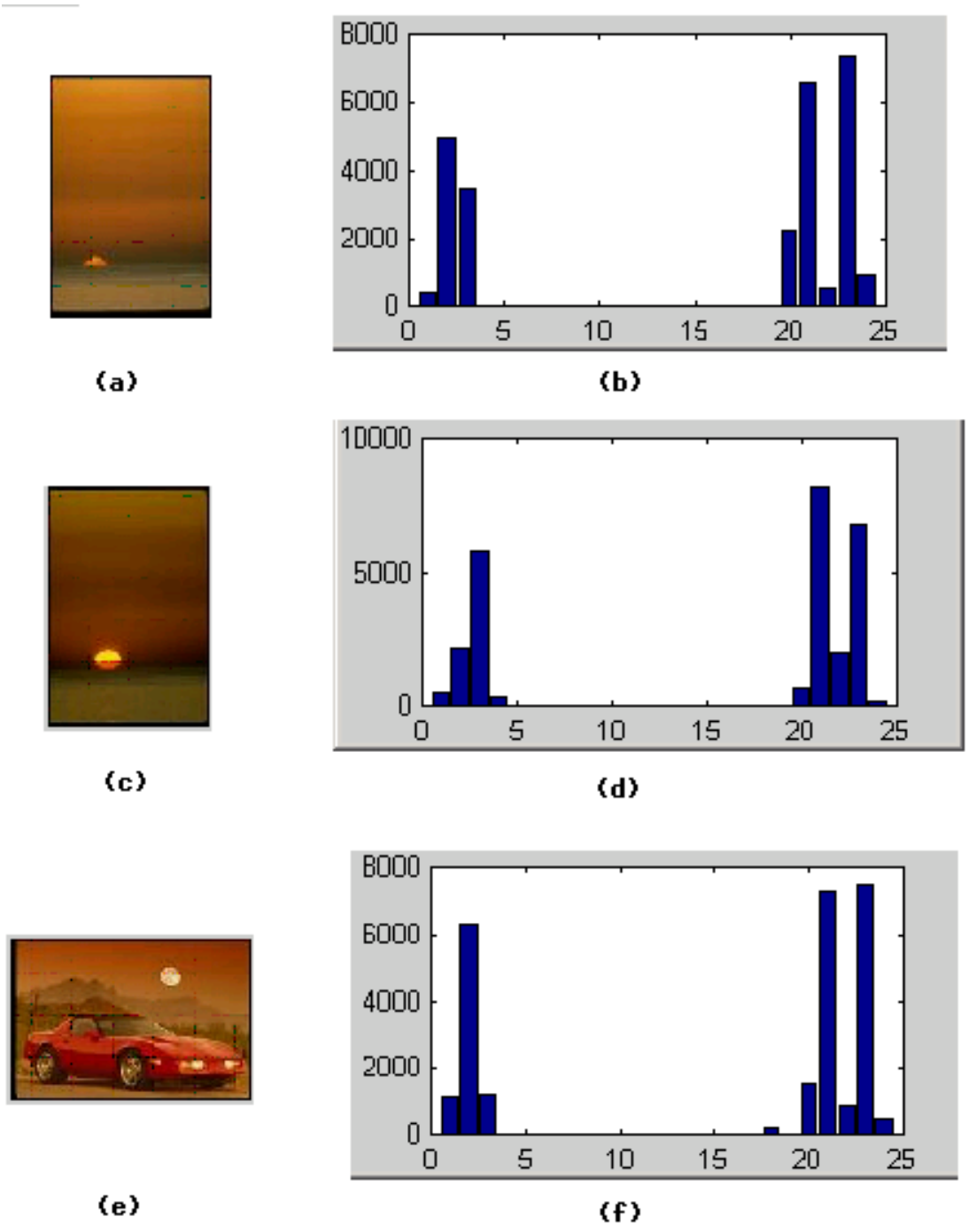


Figura 1.2: Exemplo de histogramas de imagens quantizadas para 24 pontos no espaço HSV-162. Apesar de a imagem Ia (a) e a imagem Ic (c) serem consideradas da mesma classe, Ia é quantitativamente mais semelhante à Ie (e) do que a Ic, devido Hb ser numericamente mais semelhante à Hf do que a Hd.

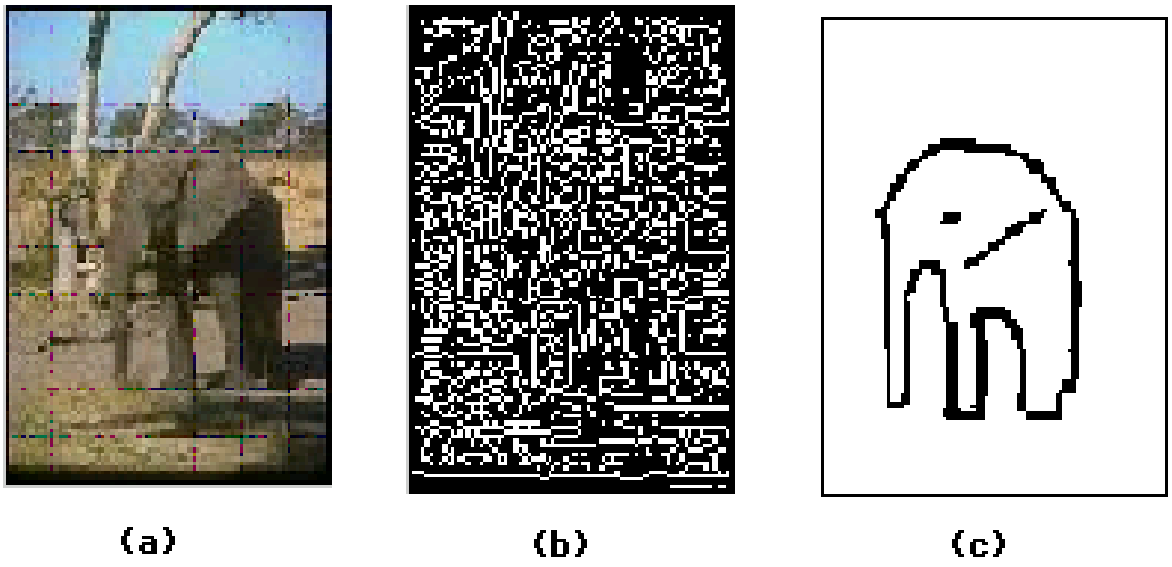


Figura 1.3: (a) Exemplo de uma imagem original colorida; (b) seu mapa de bordas; e (c) exemplo de uma esboço de um contorno externo fechado.

Características de textura

Esquemas que utilizam a textura como argumento de indexação tentam capturar informações como orientação, granularidade, densidade, etc. Para esse fim, duas abordagens podem ser destacadas: as que extraem características de matriz de co-ocorrência da imagem [Gonzalez e Woods, 1992] e as que utilizam filtros no domínio da freqüência, como os conhecidos filtros de Gabor [Drimbaream e Whelan, 2001, Manjunath e Ma, 1996]. A principal desvantagem das técnicas de textura é que, independente da abordagem que utilizam, para capturarem características em diversas orientações diferentes, precisam da aplicação de filtros direcionados quantas forem as direções requeridas. Neste trabalho, optou-se pelo uso de técnicas baseadas em informações extraídas das matrizes de co-ocorrência para capturar atributos de textura, com o intuito de fazer comparação com os modelos apresentados mais tarde.

1.1.2 O problema geral em RIBC

Baseado no que foi descrito nas seções anteriores, é natural tentar saber o porquê o uso unicamente das primitivas tradicionais (cor, forma e textura) limita o processo de RIBC, e que rumos ou soluções seria possíveis tomar para melhorar os métodos.

Geralmente, nenhum sistema de RIBC, que incorpora somente essas três características primitivas, alcança bom desempenho com relação às expectativas do usuário. Segundo [Eakins e Graham, 1999] e [Eakins, 2002], a razão desse fato é que o nível semântico geralmente requerido por usuários está bem acima do que é normalmente oferecido por sistemas que se baseiam unicamente na extração de características primitivas, como cor, forma e textura. Em [Eakins, 2002], é argumentado que atualmente as requisições do usuário são classificadas em três categorias diferentes: (a) requisições de nível 1, que são baseadas em primitivas (como cor, forma e textura); (b) requisições de nível 2, baseadas em objetos específicos (casa, carro, animais, etc.); e (c) requisições de nível 3, que são aquelas baseadas em atributos abstratos (pôr-do-sol, mar, cenas rurais, urbanas, etc.). Nesse mesmo trabalho, também argumenta-se que a maioria das requisições de usuários, hoje em dia, é de nível 2, algumas de nível 3, e quase nunca de nível 1. Entretanto, a maioria dos sistemas de RIBC oferece buscas de nível 1, poucos oferecem de nível 2 e quase nunca de nível 3.

Então, duas perguntas cabem aqui: por que a maioria desses sistemas é construída assim? E, como modelar sistemas que ofereçam buscas efetivamente de nível 2? Uma das razões do porquê pode estar no fato de que a RIBC tem suas raízes nas técnicas clássicas de análise de imagens. Mas a pergunta mais difícil de responder é, obviamente, a segunda, pois depende da modelagem e implementação de conceitos abstratos como "reconhecimento" de objetos individuais em uma cena. Como [Ullman, 1996] argumenta, a própria palavra "reconhecimento" é, por si só, ambígua; uma vez que um objeto pode ter vários significados, dependendo do observador e objetivo; além do que, para uma mesma pessoa, um determinado objeto também pode ter vários significados. Então, para serem respondidas requisições de nível 2, que incluem reconhecimento de objetos individuais, há, primeiramente, a necessidade de se definir o que significa "reconhecer". Biologicamente

falando, reconhecer envolve fatores e informações que nem sempre estão contidos somente em objetos individuais da cena, mas, na maioria das vezes, estão contidos na cena como um todo, e às vezes fora dela. Alguns desses fatores, como o contexto e o agrupamento, serão discutidos nas próximas seções. Com isso, o trabalho apresentado daqui por diante não tem a intenção de responder a esta segunda pergunta absolutamente, e sim, apenas apresentar uma alternativa de um caminho a ser seguido. Para isso, apresentam-se novas técnicas que tentam, de várias maneiras, combinar características primitivas, na tentativa de fornecer informações semânticas, ao invés da simples utilização individual das mesmas. Os resultados experimentais, mostrados no Capítulo 5, mostram que é possível melhorar o desempenho dos sistemas em RIBC combinando características sem, contudo, exigir mais informações do usuário.

1.2 O papel do contexto na Visão

Como foi abordado anteriormente, nem todos os fatores que levam ao reconhecimento são considerados na maioria dos sistemas de RIBC. Um desses fatores é o contexto, que é tradado de maneira superficial nesta seção.

Extrair características de cor, forma e textura são as estratégias escolhidas pela maioria dos trabalhos em RIBC, não importando se o objetivo é apenas recuperar as imagens "mais similares", ou simplesmente "reconhecer" um objeto ou cena. Entretanto, como foi argumentado na seção anterior, a tarefa de "reconhecer" um objeto em uma imagem, ou mesmo comparar uma imagem "semanticamente" à outra, é um processo cognitivo que depende de muitos outros fatores, difíceis (para não dizer impossíveis) de implementar. Esses fatores, no caso do sistema visual biológico, não são unicamente a cor, a forma e a textura da imagem, mas também podem ser principalmente a posição relativa dos objetos e até fatores não perceptíveis visualmente como som, cheiro, calor, pré-atenção, etc.; em outras palavras, o contexto. Contexto visual é, portanto, tudo aquilo que envolve o objeto de interesse em uma cena e que influencia na sua interpretção semântica.

Uma vez que esses são assuntos inerentemente ligados a fatores ne natureza psíquica,

é conveniente citar aqui alguns trabalhos na área da psicologia, como [Palmer, 1975] e [Chun, 2000], onde afirmam que o contexto é o fator principal para o reconhecimento de objetos. Nesses trabalhos, é argumentado o porquê, dependendo do contexto onde o objeto está inserido, o sistema visual biológico pode ser impelido a interpretar de formas diferentes. Por outro lado, citando novamente [Eakins, 2002], quando sustenta que a maioria dos usuários não está interessada em buscas por características primitivas como, cor, forma ou textura, mas sim por buscas que envolvam algum grau de relacionamento semântico entre os objetos da cena, imediatamente pensa-se também no conceito de contexto.

Falar em contexto é, por conseguinte, muito amplo. No entanto, o contexto que os trabalhos citados acima se referem é, na verdade, o contexto que envolve somente o reconhecimento de objetos em torno de outros objetos. Por exemplo, reconhecer um inseto como sendo uma abelha é mais fácil se este inseto estiver ao redor de outras abelhas. No entanto, existe um outro tipo de contexto, que não necessariamente envolve o reconhecimento de objetos, mas que também carrega informações que podem ajudar na tarefa de interpretação de uma cena. É o contexto que inclui a co-ocorrência de características primitivas, e não necessariamente de objetos semânticos. Por exemplo, a freqüência de ocorrência de uma determinada orientação ou cor pode ser utilizada para identificar a presença de determinados padrões, ou mesmo na busca de imagens semelhantes. É a modelagem e implementação desse tipo de informação contextual, mais simples, que este trabalho trata em sua grande parte.

Apesar de sua comprovada importância, poucos trabalhos utilizam informações de contexto para o reconhecimento. Pode-se argumentar, no entanto, que as técnicas clássicas e suas variantes, como histograma de cores ou orientações, também são maneiras de se ter informações contextuais, pois quando se tem um histograma de cores, por exemplo, pode-se dizer: "ocorreu 30% da cor azul e 20% da cor verde". Esse tipo de técnica é, no entanto, bastante limitada no sentido de que não leva em conta informações que co-ocorrem entre si. Em contrapartida, isso é bem diferente de se dizer: "ocorreu 30% da cor azul, concentrada na parte superior da imagem, e 20% da cor verde, concentrada na parte inferior". Sem dúvida, a segunda sentença carrega mais informação do que a primeira, restringin-

do a interpretação da cena. Isso significa que é possível combinar diversas informações de contexto de várias maneiras, com o objetivo tanto de "reconhecimento" quanto de recuperar imagens similares. No capítulo seguinte, são apresentados alguns trabalhos que procuram valer-se efetivamente desse tipo de informação, que leva direta ou indiretamente à modelagem contextual.

À luz dessas idéias, um primeiro problema será tratado neste trabalho: como aumentar a quantidade de informação relevante extraída da imagem (informações de contexto) e combiná-las de maneira adequada para serem usadas em sistemas de RIBC sem, no entanto, exigir que o usuário forneça essas informações? Neste trabalho, acredita-se que descobrir formas de modelar o contexto é o caminho a ser seguido na área.

1.3 O papel do agrupamento na visão

Um outro fator importante no processo de interpretação de objetos e cenas, já mencionado anteriormente, é o agrupamento. Essa palavra, mesmo no contexto específico da visão biológica, tem um significado muito amplo. Quando se fala em agrupar, sempre pensa-se em agrupar por "semelhança". Entretanto, diversas são as semelhanças que podem existir em uma imagem. Pode-se, por exemplo, além de semelhanças de cor, forma e textura, agrupar por proximidade de padrões verticais, horizontais, etc. Esse tipo de agrupamento segue as leis de Gestald, estabelecidas no século passado. Assim, pode-se indagar: como procede o sistema visual biológico diante de informações de semelhança? É possível modelar agrupamento? Essas questões são discutidas brevemente a seguir.

Imagine que, de repente, você perdeu toda a informação contida em sua memória e tivesse que aprender tudo de novo. Neurologistas poderiam argumentar que alguém assim não sobreviveria por si só, uma vez que seria incapaz de reaprender tudo desde o início, pois precisaria de um mínimo de informação (em forma de estímulo) para dar início hao processo de (re)aprendizado. No entanto, supondo-se que essa quantidade de informação, "mínima", exista no caso acima e que seja possível, após uma perda de memória, aprender tudo novamente, tal qual um recém-nascido ao primeiro contato com o mundo externo

ao útero materno. Qual seria, então, a primeira coisa que o sistema visual biológico faz, ao se deparar com uma cena complexa do nosso mundo, não tendo nenhum padrão, nada para comparar, com o intuito de identificar algum objeto? Teorias tradicionais, como as apresentadas por [Neiser, 1967], [Marr, 1982] e [Biederman, 1987], a respeito da percepção visual sugerem que, antes mesmo do reconhecimento de algum padrão ou objeto, mesmo em seres adultos, o sistema visual organiza ou segmenta as características de acordo com determinadas "semelhanças". [Vecera e Farah, 1997] argumentam que o agrupamento de objetos por características semelhantes é, de fato, "segmentar as características de diferentes objetos à parte, uns dos outros". Então, eles definem a segmentação de imagens como sendo "(..) o processo pelo qual o sistema visual biológico agrupa ou 'liga' localizações ou características que são parte de um mesmo objeto semântico (..)". Segmentações como esta facilitam o reconhecimento de objetos de interesse, independente de oclusões parciais. Portanto, uma das primeiras coisas que um sistema de visão biológico faz, ao se deparar com uma cena, e a despeito de não possuir nenhum padrão para comparação, é agrupar por características semelhantes; em outras palavras, segmentar regiões, e não objetos ou padrões pré-definidos.

As idéias em [Vecera e Farah, 1997] indicam uma extensão do papel do agrupamento de características (em oposição ao de objetos) não somente a uma cena ou imagem, mas a um conjunto delas. No entanto, agrupar imagens por características também é uma abordagem bastante conhecida. Neste trabalho, umas destas abordagens é oferecida com o intuito de usar seus resultados para comparação com as técnicas apresentadas como propostas. Assim, neste trabalho as bases de dados são agrupadas por características semelhantes, através do uso de técnicas que serão apresentadas mais tarde. Ao final de uma busca, ao invés de ser retornado para o usuário um conjunto de imagens ordenadas por semelhança, em resposta a uma requisição, será retornado um conjunto de grupos de imagens semelhantes. E, dentro de cada grupo, as imagens são ordenadas por similaridade decrescente.

1.4 Combinando todas as idéias pode dar certo

Quando se trata de imagens genéricas, pode-se obter resultados inesperados quando uma combinação de características distintas é usada. Um usuário, por exemplo, pode desejar obter imagens de objetos com uma determinada forma, mas especificar como "padrão de consulta" uma imagem que, além da forma, tenha cor ou textura específica, e obter resultados piores do que se fosse estabelecido como característica relevante somente a forma. Para este tipo de usuário, existem trabalhos que procuram, através de uma interação com o sistema, especificar qual ou quais características são mais relevantes para ele, e fazer a recuperação baseada nelas. Essa técnica é conhecida como "realimentação de imagens relevantes" (*relevance feedback*, no inglês) [Sclaroff et al., 1997]. Entretanto, trabalhos como os de [Greenberg, 2001] discutem o fato de que sistemas interativos podem ser cansativos e tediosos, dependendo da qualificação do usuário; ou seja, usuários especialistas têm disponibilidade de fornecer informações em um processo interativo, enquanto usuários eventuais e pouco conhecedores de uma área preferirão uma abordagem mais automática. Por esta razão, o trabalho apresentado aqui é baseado em sistemas totalmente automáticos, que utilizam tanto características primitivas quanto características espaciais, que representam informações do contexto da cena, e não somente locais. Aqui, diferentemente das seções anteriores, onde a preocupação era "como modelar informações como contexto e agrupamento", o problema é "como combinar da melhor forma possível essas informações", de modo a se obter o melhor desempenho possível a uma requisição de consulta.

Como foi argumentado previamente, diversas características podem ser extraídas de uma imagem. Três das quais (cor, forma e textura) já foram exaustivamente estudadas na literatura. Contudo, experimentos mostram que, quando usadas isoladamente ou combinadas mesmo de forma simples — como atrelar peso a cada uma —, não são eficazes na captura de informações implícitas, como informações de contexto. O fato é que, se forem fixados pesos para que determinadas características tenham sempre baixa ou alta influência no resultado final, o desempenho pode ser baixo quando, em uma consulta, a combinação de pesos requerida for muito diferente da fixada. Portanto, ora pode haver a necessidade

de se dar importâncias iguais para cada característica, ora pode haver a necessidade de importâncias relativas. Além do que, quanto maior o número de características envolvidas mais difícil será a fixação de pesos.

À luz dessas idéias, pode-se concluir que há a necessidade de se ter sistemas que combinem automaticamente diversas características heterogêneas. Assim, neste trabalho, também é proposta uma nova maneira de combinar as características envolvidas, de modo que, quando determinada característica tiver baixa prioridade na requisição, ela terá automaticamente baixa influência no resultado final. A estratégia adotada aqui baseia-se na distribuição conjunta de probabilidades dessas características, e conseqüentemente, de imagens. Essa distribuição foi implementada através de uma rede bayesiana [Ross, 1998].

1.5 Proposta da tese e contribuições

Resumo dos pontos abordados

1. Historicamente, a área de RIBC utiliza informações de cor, forma e textura como características primitivas para representar as imagens;

2. a utilização de "informações contextuais" é um meio do sistema visual biológico (no entanto, não o único) para alcançar o seu objetivo, o "reconhecimento";

3. o agrupamento de informações primitivas por semelhanças é um dos primeiros passos num processo de reconhecimento por sistemas biológicos;

4. o contexto é um dos principais fatores utilizados pelo sistema visual biológico para alcançar a sua tarefa no processo de "reconhecimento";

5. poucos trabalhos na área têm dado atenção efetiva às informações explicitadas nos itens 2, 3 e 4;

6. técnicas tradicionais que combinam características primitivas não têm obtido resultados satisfatórios.

Soluções propostas

Como hipótese de solução para representar algumas informações de contexto[3], duas metodologias foram aplicadas. A primeira usa Análise Semântica Latente (ASL), que agrega tanto informações locais quanto globais da imagem. A segunda usa um novo algoritmo, batizado de GRAS, anacrônico para *Graph Region Arrow Shot*, proposto aqui, que extrai características do relacionamento espacial entre regiões da imagem. Essas características são adicionadas ao espaço vetorial com o intuito de agregar as informações globais às primitivas locais. Essas soluções procuram atender ao item 2 acima.

Como hipótese de modelagem para o agrupamento (item 3 acima) usa-se o algoritmo de Kruskall[Balakrishnan, 1997] para encontrar a árvore geradora mínima que vai dar origem aos diversos grupos. Neste caso, em resposta a uma requisição de consulta, são ordenados grupos de imagens e não somente de imagens individuais. Nesse ponto, pode-se ratificar que essa estratégia não consta de nenhuma contribuição relevante, uma vez que classificar imagens usando algoritmos de agrupamento de dados é uma técnica já bastante conhecida.

Finalmente, atendendo ao item 4, as respostas de ordenação com base em características de cor, forma e textura, em uma base vetorial transformada com ASL, agrupada e que contenha informações espaciais, são combinadas em uma rede bayesiana.

Principais contribuições

- um novo algoritmo para extração de características de objetos e cenas, GRAS. Este algoritmo pode ser usado com imagens segmentadas de maneira simples, sem a necessidade do uso prévio de algoritmos com alto custo computacional;

- modelagem de características primitivas, que contenham informações de contexto, com o uso da ASL;

[3]Dentre os diversos tipos de informações contextuais discutidas, optou-se por tentar implementar informações primitivas. Um dos motivos, além da maior simplicidade, é o fato de ser possível fazer comparações com métodos que também usam características primitivas mas, no entanto, não levam em conta as suas ocorrências contextuais.

- um modelo de rede bayesiana para recuperação de imagens com base na combinação de diversas características primitivas;

As contribuições apresentadas aqui têm como objetivo final obter uma ordenação melhor das imagens recuperadas em uma consulta a uma determinada base, com a obtenção do maior número possível de imagens similares à consulta, contidas no banco de dados (revocação) e com maior certeza (precisão).

1.6 Organização do trabalho

A presente tese está organizada da seguinte maneira.

No Capítulo 2, são apresentados os principais trabalhos relacionados. Esses trabalhos são mostrados cronologicamente. Alguns dão ênfase ao processo de extração de características, como o trabalho proposto aqui, enquanto outros preocupam-se geralmente com a indexação. Como tanto indexação quanto extração de características estão intimamente relacionadas, ambos os tipos de trabalhos são discutidos.

O Capítulo 3 apresenta alguns conceitos básicos (independentes uns dos outros) sobre assuntos que são abordados ao longo de toda a tese.

No Capítulo 4, é apresentada a metodologia proposta, dividida em três partes principais: o algoritmo (*GRAS*), a modelagem com ASL e o modelo de rede bayesiana proposto.

O Capítulo 5 é o capítulo dos experimentos. Nele são mostrados e discutidos os resultados da execução dos métodos propostos aqui, bem como dos tradicionais. Isso é feito com o uso de duas bases de dados distintas. Uma controlada, com objetos em fundo de cena homogêneo. A utilização dessa base tem como objetivo testar os métodos propostos em sistemas que tenham a finalidade principal o "reconhecimento" de objetos e cenas, como Sistemas de Inteligência Artificial. A segunda é uma base genérica, onde as imagens foram manualmente separadas em quatorze classes distintas. Foram feitos testes com várias combinações de características diversas.

Finalmente, o Capítulo 6 apresenta as conclusões e direções futuras.

Capítulo 2

Trabalhos Relacionados

2.1 Introdução

Pesquisas em Recuperação de Imagens com Base no Conteúdo em si têm uma história relativamente recente; a maioria dos trabalhos estudados para essa tese datam de 1995 em diante. Conseqüentemente, muitas das técnicas que são usadas hoje foram adaptadas de áreas relacionadas como a clássica "reconhecimento de padrões" e "aprendizado de máquina". Com isso, nem sempre é fácil distinguir entre pesquisas em "interpretação de imagens", pura e simplesmente, e pesquisas motivadas por desejos de desenvolver sistemas de armazenamento, gerenciamento e recuperação de imagens. Também, é difícil atribuir uma determinada técnica unicamente à área de recuperação de imagens ou à análise semântica das mesmas. Estes fatos são indícios da imaturidade da área. Entretanto, recuperação de imagens com base no seu conteúdo é uma área de interesse crescente.

Dentro do campo de RIBC, como foi argüido no capítulo anterior, podem ser identificadas diversas outras áreas de atividade, onde muitas técnicas são comuns em mais de uma delas e a distinção entre as abordagens nem sempre é clara. Entretanto, a seguinte divisão é geralmente aceita [Eakins, 2002]:

- classificação automática de cenas, que, tipicamente usa métodos estatísticos;
- classificação automática de objetos, usando uma das seguintes abordagens:

técnicas baseadas em modelos de objetos previamente armazenados;

técnicas estatísticas, semelhantes àquelas usadas para classificação de cenas;

- técnicas de realimentação de informações através de usuários.

A seguir, são discutidos exemplos para cada uma dessas abordagens citadas acima.

2.2 Classificação automática de cenas

Classificação automática de cenas (em tipos gerais como "ambientes fechados", "ambientes naturais", "cidades", "animais", etc.) pode ser útil, tanto porque pode ser usada como um "filtro" antes de buscas mais específicas quanto pode ajudar na identificação de objetos pertencentes unicamente a um único contexto. Essas abordagens geralmente usam técnicas estatísticas, como histogramas de cores combinadas com análise de imagens.

Um dos primeiros sistemas desse tipo foi o Sistema IRIS [Hermes et al., 1995], o qual usa uma combinação de cor, textura e relação espacial entre as regiões para derivar uma interpretação da cena, gerando descrições do tipo "montanha", "floresta", "lago", etc. que serve de entrada para um sistema com interface baseada em texto.

Outras abordagens posteriores também seguiram a linha de tentar fazer uma análise de cena. Por exemplo, [Oliva et al., 1999] usaram filtros de Gabor sobre atributos de bordas como característica global da imagem para separar as cenas em duas classes: $artificial \times natural$ e ainda *ambiente fechado* $\times$ *aberto*. [Szummer e Picard, 1998] usaram uma combinação entre histogramas de cor, textura e Transformada Discreta de Cosseno para treinar um classificador baseado no Algoritmo do Vizinho mais Próximo para distinguir entre cenas ao *ar livre* e em *ambientes fechados*. Testes empíricos mostraram que o método obtém até 90% de precisão quando usado para classificar um conjunto de 1300 fotografias coloridas. [Lipson et al., 1997] propuseram uma abordagem diferente, baseada em uma análise semântica qualitativa, usando padrões e uma combinação de distribuição de cor para prototipar cenas como "montanhas" ou "campos". Eles divulgaram uma acurácia de 75% na classificação de fotografias de montanhas, com 12% de falsos positivos.

[Vailaya et al., 1998] desenvolveram um modelo de Classificador Bayesiano para agrupar imagens em um número de categorias, incluindo *cidades* × *paisagens* e *florestas* × *montanhas*, usando vetores gerados por quantização vetorial a partir de um espaço de momentos de cores e coeficientes de Gabor. A acurácia divulgada foi em torno de 90% para a maioria das imagens classificadas.

Um dos maiores problemas encontrados por desenvolvedores de sistemas de reconhecimento e recuperação de imagens é a medida de acurácia desses sistemas. Além do fato de ter que se definir formalmente, para cada sistema, o que é reconhecimento, ainda existem os problemas de custo computacional e dificuldades na comparação dos métodos. No caso da comparação, o problema se agrava devido existirem poucas coleções indexadas e aceitas como padrão para comparação. Todos esses problemas levam a uma certa dificuldade na medida da acurácia. Essa dificuldade é ilustrada por [Paek et al., 1999], que desenvolveram um protótipo de um sistema para classificação de fotografias de jornais em cenas de ambientes abertos e fechados, baseado em palavras-chaves de títulos de figuras e histogramas de distribuição de cores e bordas. Este sistema conseguiu 86% de acurácia e facilmente superou o método de [Szummer e Picard, 1998], o qual conseguiu apenas 74% com o mesmo conjunto de testes. Outro problema potencial está na escolha das classes às quais as imagens devem ser assinaladas, uma vez que isso sempre é uma escolha subjetiva e, portanto, limita os resultados experimentais obtidos. Uma exceção para esse problema é o chamado Classificador Hierárquico, desenvolvido por [Vailaya et al., 1998], o qual usa técnicas automáticas de agrupamento baseadas em classificação subjetiva feita por uma interface independente que gera conjunto de classes.

Um sistema que também é baseado em técnicas estatísticas, é o chamado *Query By Image Content (QBIC)* [Flickner, 1995, Niblack, 1993], da IBM, que permite busca de imagens por conteúdo usando cor, textura e forma. A estrutura de dados para organizar o espaço multidimensional dessas características é baseada no que é conhecida como R^*-*tree* [Beckmann et al., 1990]. As técnicas utilizadas pelo *QBIC* são usadas em alguns produtos comerciais (*Multimedia Manager* da *IBM*, biblioteca digital da IBM, e os produtos da série *DB2*, também da IBM). Neste sistema, para extração de características de cor, cada eixo

do espaço de cor *RGB* é quantizado em um número, K, de níveis pré-definidos, gerando um espaço de cor de K^3 células. Depois de calcular o centro de cada célula em coordenadas TMM (Transformação Matemática para Munsell), um procedimento divide o espaço em "super-células". O histograma de imagens gerado representa o número de *pixels* que pertence a cada "super-célula".

Quando uma consulta é realizada, o histograma da consulta é casado com os histogramas das imagens na base de dados. Então, a diferença, Z, entre dois histogramas é calculada com uma medida de similaridade dada por $||Z|| = Z^T AZ$, onde A é uma matriz simétrica com $A(i, j)$ representando a similaridade entre as cores i e j.

No sistema *QBIC*, as características extraídas para a textura são a granularidade, o contraste e a direcionalidade. Granularidade é usada para medir a escala da textura; o contraste mede a sua vivacidade, e depende da variância dos tons de cinza do histograma; finalmente, direcionalidade dá a direção principal da textura da imagem, e depende dos gradientes de direções.

Para a extração de características de forma, no sistema *QBIC*, assume-se que os objetos não são oclusos. São extraídos vários parâmetros de forma, que não têm exatamente o mesmo significado convencional. Esses parâmetros normalmente são a área, que é calculada como sendo o número de *pixels* contidos nas bordas das regiões; a circularidade, que é calculada como sendo o perímetro elevado ao quadrado dividido pela área; a orientação do eixo maior e a excentricidade, que são calculadas usando a matriz de covariância de segunda ordem dos *pixels* das bordas. Assim, a orientação do maior eixo é tomada como sendo a direção do maior autovetor dessa matriz, enquanto a excentricidade é a razão entre o menor autovalor e o maior. Um conjunto de momentos invariantes e um conjunto de tangentes ao redor do perímetro completam a lista de características para descrever a forma. Na fase de consulta, o cálculo da similaridade é feito em um subconjunto de características de forma selecionado pelo usuário como relevantes para expressar sua consulta. A medida de similaridade usada é dada através da Distância Euclidiana, onde são usados pesos calculados como o inverso da variância para cada característica.

Um outro sistema de recuperação de imagens com base em análise estatística de cenas

é o chamado *VisualSEEk* [Smith e Chang, 1996b], que permite a consulta através de regiões de cor e relacionamento espacial. Nele, utiliza-se uma função específica para medir a similaridade, que contém tanto informações de cor quanto informações de componentes espaciais. Para essa medida, são usadas características espaciais e de forma, como o tamanho das regiões e localização espacial. Entretanto, para derivar essas informações, há a necessidade de operações complexas: usa-se uma estrutura *Quad-Tree* ou uma *R-Tree*, no caso da consulta ser por região, e um esquema que utiliza *strings* 2D para representar relações espaciais em imagens com múltiplas regiões. A principal desvantagem dessa abordagem é que, para cada tipo de consulta realizada, há a necessidade de se utilizar um desses tipos diferentes de soluções, não integrando, no entanto, um único método para indexação.

O VIPER [Ooi et al., 1998] é outro sistema de recuperação de imagens que emprega tanto cor quanto informação espacial para facilitar o processo de recuperação. Primeiro, um conjunto de cores dominantes é extraído. Em seguida, são extraídas informações espaciais a partir das regiões delimitadas por essas cores dominantes. Então, nesse sistema, duas imagens são similares em termos de cor e informações espaciais se elas possuem características semelhantes que pertençam ao mesmo espaço vetorial.

O sistema SCARLET (*System for Content-based imAge Retrieval using waveLET*) [Lee e Kim, 2001] apresenta um método para extração de características das bordas através de transformação *wavelet*. O método obteve resultados semelhantes aos obtidos pelo *QBIC*. Entretanto, a ênfase dada neste sistema está na indexação; ou seja, no acesso à base de dados. Para isso, apresentou-se uma nova estratégia de busca que utiliza espaço multidimensional, batizada de *SPY-TEC* (*Spherical Pyramid-Technique*). Esta técnica particiona o espaço d-dimensional, primeiro em espaços piramidais bidimensionais, e depois particiona as pirâmides em pequenos pedaços. Esta partição transforma o espaço d-dimensional em um espaço unidimensional. Assim, é possível o uso de uma estrutura de árvore B^+-*tree* para gerenciar o espaço unidimensional, que acelera o processo de busca.

O trabalho apresentado por [Hirata e Kato, 1992] apresenta um método para recuperação de imagens através de exemplos visuais. Neste sistema, chamado de "consulta através de exemplos", são extraídas bordas das imagens de consulta e comparadas com

as imagens do banco de dados através de um processo complexo que necessita realizar deslocamentos e deformações nas imagens.

[Jacob et al., 1995] propuseram um método para recuperação de imagens, que utiliza informação espacial e extração de características utilizando coeficientes dominantes de *wavelet*. Esse método procura melhorar a eficiência na busca de similaridade e extração de características usando transformada de *wavelet*. No entanto, a base de dados utilizada, por ser pequena, não é muito representativa.

No recente trabalho apresentado por [Zhou e Huang, 2001], foi mostrado um algoritmo que extrai características exclusivamente do mapa de bordas da imagem original. O algoritmo percorre o mapa e, durante a varredura, extrai características espaciais consideradas globais da imagem. Esses resultados mostraram-se eficientes com relação à precisão. A análise dos resultados indica que é possível se obter bom desempenho, mesmo a partir de características exclusivamente espaciais[1].

Sistemas como os discutidos nesta seção fornecem um certo grau de classificação e recuperação semântica de imagens, permitindo atribuir rótulos como "montanhas", "praias" ou "cidades" às imagens. Até o momento, eles tendem a representar uma abordagem — embora que geral para recuperação semântica — que faz apenas uma classificação do tipo X ou *não-X*, onde X deve ser uma cena de determinada classe. Muitos dos resultados que obtiveram uma boa acurácia são baseados nessa abordagem. Isto é devido ao fato de ser discutível que sistemas como esses exibem "comportamento inteligente". Na verdade, eles incorporam muito pouco conhecimento do que é chamado na literatura de "visão de alto nível" [Ullman, 1996]. Em particular, poucos têm a habilidade de continuar a "aprender" durante o estado operacional. Uma exceção a esta regra é o sistema descrito por [Vailaya e Jain, 1999]. Este sistema gera um espaço vetorial através de um algoritmo de quantização e atribui pesos aos pontos deste espaço. A cada novo dado acrescentado ao espaço, os pesos são recalculados. Assim, o sistema é sempre atualizado com relação aos

[1] O algoritmo GRAS, proposto aqui, segue essa linha de pensamento. No entanto, ele difere do trabalho de[Zhou e Huang, 2001], principalmente em dois pontos. Primeiro que a varredura é feita sobre um grafo, o qual representa as regiões da imagem. Segundo, as características extraídas por ele são combinadas de forma adequada com outras características, como cor, forma e textura.

dados novos, simulando um certo grau de "aprendizado".

2.3 Classificação automática de objetos

Claramente, reconhecimento automático de objetos é uma tarefa importante para recuperação semântica de imagens. A identificação de um dado tipo de objeto em uma cena é útil tanto como um objetivo em si quanto como um passo intermediário na interpretação de cenas mais complexas. [Vailaya e Jain, 2000] propuseram, por exemplo, um esquema para indexar imagens que combina identificação de cenas usando características globais e detecção de objetos usando características locais. As duas principais abordagens usadas para reconhecimento de objetos podem ser descritas respectivamente como aquelas baseadas em "modelos globais" e aquelas baseadas em "técnicas estatísticas". Entretanto, não é fácil estabelecer uma distinção clara entre as duas. A seguir, serão descritas brevemente cada uma delas.

2.3.1 Técnicas baseadas em modelos

Uma das abordagens mais efetivas para reconhecimento de objetos em uma imagem pode ser "especificar um modelo para cada objeto diferente". Desta maneira, é necessário que o software usado para analisar a imagem seja capaz de identificar o modelo de objeto dentro da cena. Uma das primeiras implementações desse princípio foi o Sistema ACRONYM [Brooks, 1983], o qual usava modelos de contornos genéricos para identificar e localizar instâncias de determinados objetos em fotografias aéreas. Após um processo inicial de detecção de bodas, descritores de possíveis objetos de interesse eram produzidos para identificação dos mesmos. Um conjunto de "regras de produção" era então usado para inferir a presença de determinados tipos de aviões.

Uma estratégia similar foi usada por [Matsuyama e Hwang, 1990], os quais desenvolveram o sistema SIGMA, onde era usado um controle hierárquico para reconhecer objetos de imagens aéreas através de uma estratégia de múltiplos níveis de detalhes. O SIGMA

possuía quatro módulos: um para extração de características de baixo nível (primitivas); um para selecionar o objeto a ser identificado; um para o reconhecimento propriamente dito; e um para interagir com o usuário. Esses módulos são interconectados através de uma estrutura *top-down* e *bottom-up*.

Outro sistema dessa mesma linha, é o chamado sistema SCHEMA, desenvolvido por [Draper et al., 1989], o qual também usa uma abordagem baseada em modelos para interpretação de cenas. Rotinas de baixo nível são aplicadas às imagens para extrair descritores, os quais são chamados de *tokens*, e posteriormente são organizados em estruturas abstratas que podem ser associadas a instâncias de objetos. Os milhares de *tokens* gerados podem ser agrupados de uma maneira combinatória. Desta forma, o "conhecimento" no SCHEMA não está limitado à descrição de objetos, puramente; ele contém informações a respeito de como cada objeto pode ser reconhecido. Assim, são formadas hipóteses a respeito dos objetos na cena. O SCHEMA produz interpretação de imagens baseada em cenas bidimensionais.

Sistemas como ACRONYM não foram desenvolvidos com o objetivo de nenhuma aplicação específica. O principal objetivo de seus idealizadores era mostrar a plausibilidade das técnicas que se baseiam em modelos de objetos para interpretação de cenas. Posteriormente, pesquisas adaptaram as técnicas testadas no ACRONYM para o domínio específico de RIBC. Um dos sistemas deste tipo é o PICTION [Srihari, 1995], o qual identifica faces humanas em cenas naturais casando "faces candidatas" geradas por técnicas que utilizam detecção de bordas a partir dos contornos da face, como os seus lados, cabelo, etc. Uma técnica mais sofisticada deste tipo baseia-se na extração do chamado "objeto composto" [Durand et al., 1999]. Efetivamente, trata-se de um sistema especialista para reconhecimento e caracterização de objetos compostos de regiões conectadas. Cada objeto composto tem que ser definido como um modelo consistindo de um ou mais componentes. Imagens de pessoas (com roupa), por exemplo, podem ser modeladas como um arranjo específico de primitivas como face, cabelo, blusa ou chapéu. Esse sistema separa as regiões usando meios convencionais e então tenta caracterizar essas regiões comparando as composições com os modelos.

Talvez a melhor técnica nesta área foi a apresentada por [Forsyth et al., 1997]. Sua abordagem é baseada em um modelo para cada classe de objeto a ser reconhecido e no uso desses modelos para encontrar evidências dessas classes na imagem. Evidências podem incluir características da região em si (como cor, forma ou textura) ou informação contextual como sua posição e o tipo de fundo da imagem. A classificação de objetos é um processo em três estágios: (a) segmentação de imagens em regiões coerentes, usando uma combinação de cor, forma e textura; (b) combinação de características de cor, forma e textura para identificar possíveis descritores de cada região (por exemplo, uma parte do corpo de uma pessoa); (c) classificação propriamente dita, utilizando esses descritores. O método foi aplicado com algum sucesso à identificação de uma determinada faixa de classes de objetos, incluindo pessoas, cavalos e árvores, embora a acurácia de recuperação do sistema seja relativamente modesta até o momento (em torno de 15% de revocação e 66% de precisão [2] para a classe cavalo, por exemplo).

Sistemas baseados em modelos de objetos são talvez mais "inteligentes" do que os baseados em estatísticas para classificação de cenas, descritos na Seção 2.2. Certamente, eles fazem extenso uso de técnicas de Inteligência Artificial. Seus modelos são freqüentemente implementados como uma base de conhecimento para o sistema, pois são relativamente capazes de "discernir" acerca da natureza dos objetos envolvidos usando uma base de modelos que é geralmente esparsa e heterogênea. Neles, a análise semântica geralmente usa métodos heurísticos guiados por conhecimento com abstração relativamente alta — os modelos de objetos propriamente ditos —. Mas, de uma maneira geral, sofrem com dois tipos de limitações. Primeiro, a natureza dos problemas que tentam resolver — compartilhados por diversos sistemas especialistas — estão, acima de tudo, relacionadas com o fato de que a base de conhecimento utilizada é extremamente dependente do domínio da aplicação e, conseqüentemente, só podem manipular um conjunto restrito de classes de objetos. Segundo, o conhecimento acumulado previamente nos modelos vem de pessoas

[2] Precisão é definida como a porcentagem de objetos recuperados relevantes para a consulta; e revocação é definida como sendo a porcentagem de objetos em toda a base de dados, relevantes para a consulta. Ver Seção 5.4 para uma definição mais completa com exemplos de precisão e revocação

que alimentam a base de modelos, logo, não possuem nenhum mecanismo para aumentar o conhecimento da base. Também, como os classificadores de cenas, não está claro como podem se adaptar para resolver situações que não são meramente distinguir entre *X/não-X*, uma vez que em situações reais, é possível se deparar com centenas, talvez milhares, de diferentes tipos de objetos, e às vezes dentro de uma mesma classe. Não importa quão rico seus modelos de objetos possam ser, sempre haverá pouco sucesso na tentativa de reconhecer alguns exemplos de novos tipos ou vistas de objetos [Rosch et al., 1976].

2.3.2 Técnicas estatísticas

Uma abordagem considerada simples para interpretação de imagens, a qual não necessita da construção de nenhum modelo de objeto de alto-nível, é o uso de técnicas estatísticas — freqüentemente, muito similares àquelas usadas em classificação de cenas — para atribuir rótulos individuais às regiões da imagem. Um bom exemplo dessa abordagem é o trabalho de [Campbell et al., 1997], que usou uma combinação de características de cor e textura para treinar uma rede RBF (*Radical Basis Function*) para distinguir entre 11 tipos diferentes de regiões em uma cena, incluindo céu, vegetação, estradas, edifícios, cenas rurais e objetos em movimento — tipicamente carros. Foi divulgado 80% de acurácia quando da classificação de 3700 regiões em 350 imagens. [Vailaya e Jain, 2000] usaram sua técnica de classificação de imagens [Vailaya et al., 1998], que se baseia em cor e textura, para reconhecimento de imagens de céu e vegetação com resultados preliminares animadores. Eles trabalham para estender a técnica na classificação de uma variedade maior de tipos de imagens.

Trabalhos similares foram desenvolvidos em uma série de laboratórios; entre eles, pode-se destacar [Martinez e Serra, 1999] que usaram análise discriminante baseada em vetores de características, derivada a partir de Análise em Componentes Principais (PCA) em imagens convoluídas com uma gaussiana, para classificar imagens em uma variedade de categorias, incluindo animais, pessoas, carros e casas. Pouca informação foi fornecida a respeito das medidas de desempenho do método. [Belongie et al., 1998] desenvolveram uma

representação chamada de *blobworld* para regiões de imagens, baseada em segmentação de cor e textura, usando um algoritmo de maximização do valor esperado. Embora não tenham sugerido que sua técnica ofereça recuperação semântica, eles mostram que pode ser usada para recuperar imagens de objetos como tigres e aeronaves. [Leung e Malik, 1999] desenvolveram um método para identificar materiais dentro de regiões de textura de imagens (como couro, cortiça, plástico, etc.) usando microestruturas conhecidas como *3D-textons*, derivadas de primitivas medidas previamente. A um nível mais especializado ainda, [Bregler e Malik, 1997] usaram algumas técnicas de medidas de textura e treinaram um classificador hierárquico capaz de distinguir entre cinco tipos de veículos diferentes. [Schneiderman e Kanade, 1998] mostraram que um classificador Bayesiano, baseado em vetores derivados com PCA da intensidade dos pixels em sub-regiões de imagens quantizadas em três níveis de resoluções, pode detectar corretamente cerca de 90% de faces em uma coleção de imagens com uma porcentagem de falsos positivos menor do que 12%. Este modelo mostrou-se superior a classificadores de faces anteriores em redes de *back-propagation* [Rowley et al., 1998].

Uma idéia com grande potencial aplicativo é o uso de classes geradas automaticamente para reconhecimento de objetos, proposta por [Schiele e Crowley, 1997]. O objetivo era resolver o problema da grande variabilidade do número de vistas de objetos como, por exemplo, "cadeiras". Então, imagens de cadeiras podiam ser detectadas usando as técnicas previamente desenvolvidas pelos autores para reconhecimento de objetos usando histogramas multidimensionais [Schiele e Crowley, 1996]. Os autores não forneceram nenhuma evidência convincente de que seus conceitos podem ser estendidos a uma base de dados que não seja a experimentada por eles. Não está claro como subclasses homogêneas de objetos, como cadeiras, podem ser identificadas; nem está claro que histogramas de campo multidimensionais são úteis para identificar diferentes instâncias de uma dada classe de objetos visualmente similares.

Uma outra técnica para reconhecimento de objetos — como, por exemplo, frutos — ou tipos de materiais — como areia — em uma imagem, é o método de [Buijs e Lew, 1999], que descreve cenas a partir de características primitivas, identificando tanto exemplos negativos

quanto positivos de imagens. Posteriormente, essas características são usadas para treinar um classificador de distância mínima.

Abordagens estatísticas têm a vantagem de não requerer a construção de um domínio de modelos específicos para cada tipo de objeto a ser reconhecido, embora eles, obviamente, sofram da falta de conhecimento de alto nível acerca do domínio. Isto porque baseiam-se totalmente em associações estatísticas entre características de baixo nível de imagens (quantificáveis), e aprendendo, em muitos casos, a partir de um conjunto de treinamento de poucas centenas de exemplos, quando muito.

Quando se julga o critério semântico requerido pelo usuário, como abordado na Seção 1.1.2, as técnicas estatísticas podem parecer menos "inteligentes" do que as baseadas em modelos de objetos, devido à falta de informações de alto-nível — ou mesmo, em muitos casos, da habilidade de lidar com informações de incerteza —. Entretanto, isso não torna essas técnicas menos útil.

2.4 Técnicas de realimentação de informações relevantes

O problema de realizar efetivamente buscas que envolvam algum conhecimento semântico puramente automático tem levado os pesquisadores a investigar métodos que possuam algum grau de envolvimento humano durante a operação dos sistemas. Esses métodos são conhecidos como Realimentação de Relevantes (*Relevance Feedback*, do inglês), os quais constam basicamente de exibir um resultado de uma busca a cada requisição do usuário, e "pedir" que seja indicado quais imagens são mais relevantes. De posse dessa informação, os sistemas são realimentados e calculam novos pesos para as medidas de similaridade, retornando uma nova busca.

Bons exemplos do uso dessa técnica podem ser encontrados em [Fournier et al., 2001, Salton e Buckley, 1990, Nastar et al., 1998].

Entretanto, já foi argumentado na Seção 1.4 que sistemas como esses tendem a ser

tediosos para usuários não muito especializados. Outro ponto a se destacar nessas técnicas é o fato de que muitas das abordagens utilizadas são comuns — ou exatamente as mesmas — às abordagens usadas nas buscas baseadas em modelos de objetos e métodos estatísticos. Entretanto, existe uma diferença crucial entre elas: os métodos que utilizam realimentação de relevantes podem continuar a aprender à medida que interagem com o usuário final na etapa *online* do processo. Isso constitui uma vantagem sobre outros métodos.

A abordagem com realimentação de relevantes, por não ser objeto de assunto dessa tese, não será discutida neste trabalho, embora possua resultados até agora comprovadamente melhores.

2.5 Conclusão

À luz dos trabalhos apresentados neste capítulo, pode-se observar claramente que cada um tenta dar uma contribuição em uma etapa diferente dos sistemas (pré-processamento, extração de características, indexação e busca). No entanto, a maioria deles apresenta um sistema completo de RIBC para validar seus métodos. Atualmente, as etapas que têm recebido mais atenção são, de longe, as de extração de características, com o intuito de melhorar a medida de similaridade; e a indexação, com o intuito de melhorar o acesso quando a base de dados é extensa. Tem-se, no entanto, um número relativamente menor de trabalhos que procuram dar ênfase à etapa de pré-processamento[3], ou à melhora da ordenação dos dados.

O trabalho que está sendo apresentado aqui pode ser incluído tanto entre aqueles que pretendem dar uma contribuição na etapa de extração de características quanto entre aqueles que pretendem melhorar a ordenação das imagens recuperadas. Essas duas etapas são realizadas através do algoritmo *GRAS* e do modelo bayesiano de recuperação, respectivamente.

[3]Talvez porque esta tarefa seja exclusivamente de Processamento Digital de Imagens e não de RIBC.

Capítulo 3

Conceitos básicos

Neste capítulo, para cada uma das propostas de solução que foram apresentadas no Capítulo 1, são apresentados conceitos básicos. Assim, a Seção 3.1 mostra o significado dos termos objeto e cena. Embora esses termos já tenham sido largamente usados, de maneira informal, nos Capítulos 1 e 2, é importante definí-los dentro do âmbito da tese. A Seção 3.2 apresenta o conceito de vistas de objetos, que será utilizado principalmente no capítulo de experimentos. Documento, coleção e consulta são termos inerentemente ligados à área de Recuperação de Informação em geral, e são apresentados na Seção 3.3. A Seção 3.5 mostra os termos, a nomenclatura e o desenvolvimento da Teoria da Decomposição em Valor Singular, que será usada dentro da teoria da Análise Semântica Latente, um dos métodos propostos na tese. A Seção 3.6 mostra como a teoria das redes bayesisnas pode ser aplicada à RIBC. Finalmente, a Seção 3.7 apresenta uma justificativa para a medida de similaridade usada neste trabalho.

3.1 Objetos e cenas

Esta seção supõe as imagens em tons de cinza. Neste trabalho, consideram-se todos os *pixels* de uma imagem com rótulo igual a 0 como sendo parte do fundo (*background*, do inglês). Os valores de rótulos de *pixels* que não são considerados parte do fundo fazem

parte de algum objeto, dependendo do seu valor de rótulo. Assim, todos os *pixels* com rótulos iguais a 1, conectados ou não, pertencem a um mesmo objeto; igual a 2, a outro objeto; e assim por diante, conforme ilustrado na Fig. 3.1.

0	0	0	0	1	0	1	0	0
0	1	1	0	0	0	0	0	0
0	1	1	1	0	0	0	0	0
0	1	1	0	2	2	2	0	0
0	0	0	0	2	2	2	0	0
0	0	0	0	2	0	0	0	0
3	3	3	0	0	0	0	0	0
3	3	0	0	3	0	0	0	0

Figura 3.1: Exemplo de estrutura de dados usada neste trabalho: *pixels* com valores de rótulos iguais a 0 representam o fundo; *pixels* com valores de rótulos iguais a 1, 2 e 3, representam regiões de objetos diferentes. O objeto 2 é formado por apenas uma região, e os objetos 1 e 3 são formados por três e duas regiões, respectivamente.

Nesta tese, os experimentos foram conduzidos em cima de dois tipos distintos de imagens: imagens com um único objeto e imagens de classes de cenas naturais, onde são consideradas as regiões e não somente os objetos. Um exemplo real de uma imagem de objeto, usada neste trabalho, é mostrado na Fig. 3.2; e um exemplo de imagem de uma cena natural é mostrado na Fig. 3.3.

3.2 Vista de objetos

Neste trabalho, o conceito de "vista de objetos" é menos formal do que o conceito de objetos e cenas, e está diretamente ligado ao conceito de perspectiva de uma imagem, onde cada perspectiva diferente chama-se "vista do objeto". Intuitivamente, o conjunto de vistas representa o mesmo objeto. A Fig. 3.4 ilustra este conceito. Nela, 16 imagens estão distribuídas em uma matriz 4 x 4 de imagens de objetos. Cada linha representa 4

Figura 3.2: Exemplo de uma imagem de um objeto usada neste trabalho. Todas as imagens de objetos usadas possuem fundo homogêneo (rótulo = 0) com apenas um objeto isolado.

vistas (de acordo com o conceito de vista definido aqui) diferentes de um tipo de objeto.

3.3 Documentos, coleção e consulta

Em recuperação de informação textual, a palavra "documento" significa um elemento recuperado, podendo ser uma página *html*, um *link*, etc. Uma coleção, normalmente, é um conjunto de documentos sobre a qual são feitas as buscas.

O equivalente a documentos, no caso de RIBC, são as imagens retornadas como resposta, e a coleção é o conjunto de imagens processadas e indexadas na base de dados da imagem.

A palavra consulta é utilizada aqui para se referenciar à imagem apresentada como padrão de comparação, muitas vezes chamada de imagem-exemplo. Pode ser tanto uma imagem real quanto um esboço de uma imagem de um objeto que se deseja recuperar ou um conjunto de imagens.

Figura 3.3: Exemplo de uma cena natural (paisagem) usada neste trabalho. Todas as cenas são de imagens naturais de diversas classes.

3.4 Informações de matrizes de co-ocorrência

Entre os métodos clássicos implementados para efeito de comparação com os métodos propostos está a técnica de extrair informações de texturas a partir da matriz de co-ocorrência da imagem [Gonzalez e Woods, 1992].

Seja P um operador de posição e A uma matriz $k \times k$ cujo elemento a_{ij} é o número de vezes em que um *pixel* com nível de cinza z_i ocorre (na posição especificada por P) em relação ao ponto com nível de cinza z_j, com $1 \leq i, j \leq k$. Por exemplo, considere uma imagem com três níveis de cinza, $z_1 = 0$, $z_2 = 1$ e $z_3 = 2$, da seguinte maneira:

$$\begin{matrix} 0 & 0 & 0 & 1 & 2 \\ 1 & 1 & 0 & 1 & 1 \\ 2 & 2 & 0 & 0 & 0 \\ 1 & 1 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{matrix}$$

Figura 3.4: Uma matriz 4 *x* 4 de imagens de objetos, tomadas como exemplos usados neste trabalho. Cada linha da matriz representa um objeto com 4 vistas distintas.

Se o operador de posição P for definido como sendo "um *pixel* à direita e um *pixel* a baixo", a seguinte matriz 3 x 3 é calculada:

$$A = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 3 & 2 \\ 0 & 2 & 0 \end{bmatrix}$$

onde, por exemplo, a_{11} (topo à esquerda) é o número de vezes em que um ponto com nível de cinza $z_1 = 0$ aparece na posição abaixo e à direita de um *pixel* com o mesmo nível ce cinza, e a_{13} (topo à direita) é o número de vezes em que um ponto com nível $z_1 = 0$ aparece um *pixel* abaixo e um *pixel* à direita de um *pixel* com nível de cinza $z_3 = 2$. O tamanho de A é determinado estritamente pelo número de níveis de cinza possíveis na imagem de entrada. Então, normalmente é necessária uma redução do número de níveis de cinza original de maneira a tornar A menor.

Seja n o número total de pares de pontos na imagem que satisfaz P (no exemplo anterior, $n = 16$). Se a matriz C é formada dividindo cada elemento de A por n, então c_{ij} é uma estimativa da distribuição de probabilidade de que um par de pontos satisfazendo P tenha valor (z_i, z_j). Essa matriz C é chamada de "uma matriz de co-ocorrência dos níveis de cinza". Devido ao fato de C depender de P, a presença de um determinado padrão de textura pode ser detectado escolhendo-se apropriadamente P. Por exemplo, o operador usado no exemplo anterior é sensível às bandas de direções -45^o. De maneira genérica, o problema consta em analisar uma dada matriz C com o objetivo de se caracterizar a região da textura sobre a qual C é computada. Um conjunto de descritores úteis para essa finalidade é:

- máxima probabilidade: $max(c_{ij})$;
- momento de ordem k: $\sum_i \sum_j (i-j)^k c_{ij}$;
- momento inverso de ordem k: $\sum_i \sum_j c_{ij}/(i-j)^k$, para $i \neq j$;
- entropia: $-\sum_i \sum_j c_{ij} \log c_{ij}$;
- uniformidade: $\sum_i \sum_j c_{ij}^2$.

A idéia básica é tentar caracterizar o conteúdo de C através desses descritores. Por exemplo, a primeira propriedade dá uma indicação da forte resposta de P (no exemplo anterior). O segundo descritor é relativamente baixo quando os maiores valores de C estão próximos da diagonal principal, devido ao fato de a diferença $(i-j)$ ser pequena neste caso. O terceiro descritor fornece o efeito oposto. O quarto descritor é uma medida da dispersão, tendo seus maiores valores quando C tende a ser homogêneo. Finalmente, o quinto descritor tem seus valores baixos quando os valores c_{ij} tendem a ser iguais.

Esses valores foram usados por serem, entre outros possíveis, os mais sugeridos na literatura [Gonzalez e Woods, 1992].

O valor de K, usado acima, foi 2, uma vez que [Gonzalez e Woods, 1992] sugere que esse valor é suficiente para representar a maioria das informações; acima dele, os efeitos são imperceptíveis.

3.5 Decomposição em valores singulares

A teoria da Decomposição em Valores Singulares[1] tem diversas aplicações em inúmeras áreas. Nesta tese, será aplicada à Análise Semântica Latente, uma das técnicas propostas para melhorar as buscas em RIBC. Nesta seção, essa teoria é descrita brevemente.

Seja A uma matriz $m \times k$ de números reais. Então existe uma matriz ortogonal U e uma matriz ortogonal V de maneira que:

$$A = USV^t,$$

onde a matriz S, $m \times k$, possui entradas $(i, j) = \lambda_i \geq 0$ para $i = 1, 2, ..., min(m, k)$ e as outras entradas iguais a zero. As constantes positivas λ_i são chamadas *os valores singulares* de A. As primeiras r colunas de V são chamadas de *os vetores singulares à direita* e as primeiras r colunas de U são chamadas de *os vetores singulares à esquerda*.

Um tratamento a respeito dessa teoria pode ser encontrado em [Golub e Loan, 1989], e um exemplo numérico simples pode ser encontrado em [Johnson e Wichern, 1998].

3.6 As redes bayesianas e sua utilização em RIBC

As redes bayesianas são um ferramental gráfico de representação das dependências entre variáveis de uma distribuição conjunta de probabilidade. Essa distribuição é modelada por um grafo direcionado acíclico, no qual os nodos representam as variáveis aleatórias da distribuição e os arcos representam os relacionamentos entre essas variáveis.

O princípio fundamental é que as independências conhecidas entre as variáveis aleatórias do domínio são declaradas explicitamente na rede e que a distribuição conjunta de probabilidade pode ser inferida a partir dessas independências. Os relacionamentos entre os nodos, indicados pelos arcos direcionados do grafo, representam dependências causais entre

[1] A decomposição em valores singulares foi descoberta independentemente por [Beltrami, 1873] e [Jordan, 1874]. [Schmidt, 1907] estudou uma decomposição análoga em seu trabalho sobre Equações Integrais.

variáveis do domínio. A intensidade dessas dependências é expressada por probabilidades condicionais associadas aos arcos do grafo.

Sejam X e Y variáveis aleatórias representadas por nodos da rede bayesiana, e sejam x e y seus respectivos valores. Se o valor da variável X é causa direta de influência no valor da variável Y, um arco direcionado de X para Y é inserido no grafo (X é dito pai de Y). A intensidade do relacionamento entre X e Y é expressada pela probabilidade condicional $P(y|x)$. Diz-se, probabilidade de $Y = y$, caso $X = x$.

Seja B uma rede bayesiana, seja Y uma variável aleatória em B, seja y o valor de Y, seja $\mathbf{P_Y}$ o conjunto de variáveis pais de Y e seja p_y o conjunto dos valores de $\mathbf{P_Y}$, como indicado na Fig. 3.5.

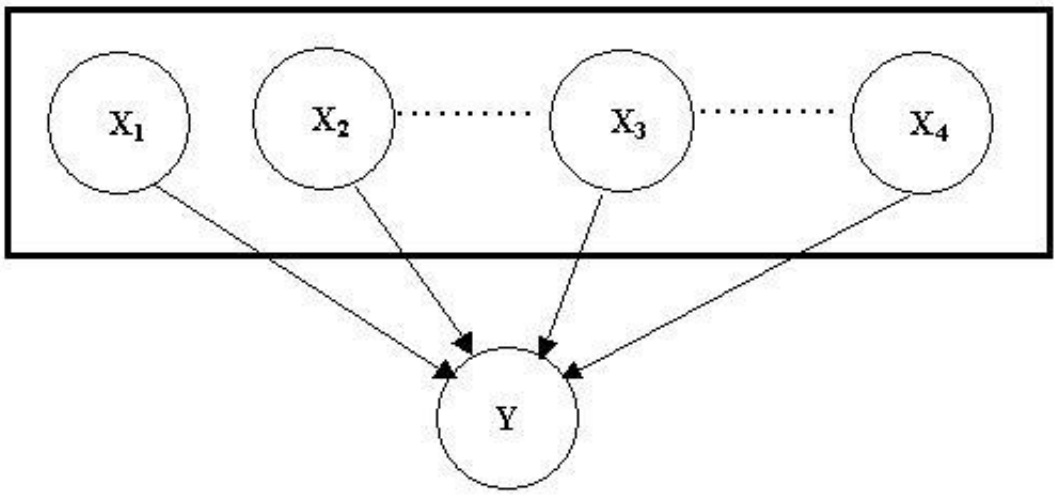


Figura 3.5: Exemplo de rede bayesiana. As variáveis X_i são causas de Y, e as probabilidades de suas influências são dadas explicitamente na rede.

A influência de $\mathbf{P_Y}$ sobre Y pode ser especificada por uma função F que satisfaça:

$$\sum_{\forall y} F(y, p_y) = 1 \qquad 0 \leq F(y, p_y) \leq 1 \tag{3.1}$$

A função $F(y, p_y)$ fornece uma quantificação da probabilidade condicional $P(y|p_y)$. Esta

especificação representa a distribuição conjunta de probabilidade para os nodos da rede B [Pearl, 1988].

Para exemplificar, considere a Fig. 3.6, que representa a distribuição conjunta de probabilidade $P(x_1, x_2, x_3, x_4, x_5)$ para as variáveis aleatórias $\{X_1, X_2, X_3, X_4, X_5\}$, onde $\{x_1, x_2, x_3, x_4, x_5\}$ são seus respectivos valores. O nodo X_1 é denominado raiz da rede (nodo sem pai) e é o pai de X_2 e X_3. A probabilidade $P(x_1)$, associada com o valor x_1 do nodo raiz X_1, é denominada probabilidade *a priori* e é utilizada para representar o conhecimento prévio sobre o domínio modelado. Dado o valor da variável X_1, as variáveis X_2 e X_3 são independentes. Dados os valores das variáveis X_2 e X_3, as variáveis X_4 e X_5 são independentes. Devido a essas independências, a distribuição conjunta de probabilidade $P(x_1, x_2, x_3, x_4, x_5)$ pode ser calculada por:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)\ P(x_2|x_1)P(x_3|x_1)\ P(x_4|x_2, x_3)\ P(x_5|x_3) \tag{3.2}$$

3.7 Medidas de similaridade

Os sistemas de recuperação de informação de texto adotam palavras-chave para indexar os documentos e também para traduzir a necessidade de informação do usuário a uma consulta. A representação de documentos e consultas por palavras-chaves é uma simplificação, pois parte da semântica dos documentos e da necessidade de informação é perdida quando é feita a transformação em um conjunto de palavras-chave. O mesmo ocorre quando a recuperação é de informações de imagens, quando se transforma o conteúdo dela em um vetor de características, pois parte do conteúdo semântico da mesma também é perdido.

Outra simplificação que ocorre nos dois sistemas também faz com que informações sejam perdidas nessas transformações: as ocorrências de palavras-chave (em recuperação textual) e características (em recuperação de imagens) são consideradas independentes; ou seja, conhecer alguma informação sobre a palavra ou característica i nada implica sobre

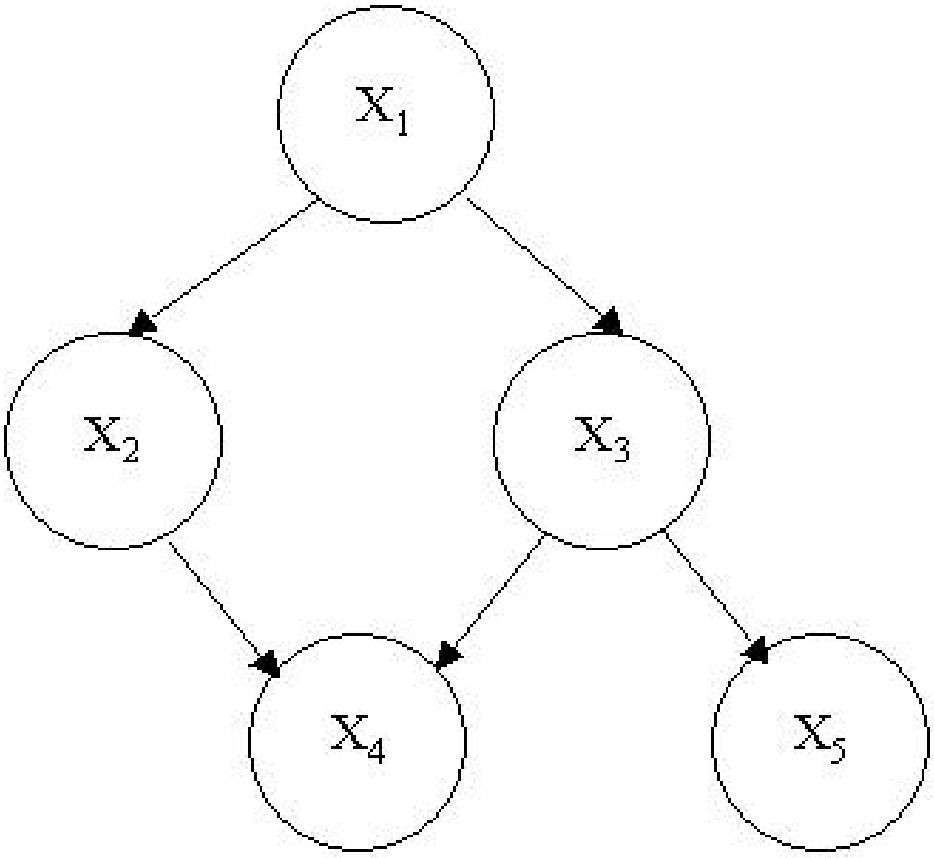


Figura 3.6: Rede Bayesiana com independências declaradas.

seu vizinho $i + 1$.

Entretanto, quando se tem um espaço vetorial e deseja-se medir a distância entre 2 vetores — e conseqüentemente entre 2 imagens — pode-se utilizar diversas métricas. Algumas delas são descritas abaixo:

3.7.1 Distâncias clássicas

Distâncias histogrâmicas são fáceis de computar com complexidade *O(n)*. Esse é um dos motivos pelos quais elas são largamente usadas em recuperação de informação. Para exemplificar algumas das mais usadas, primeiramente, considere duas distribuições, H_0 e H_1, com tamanho n cada uma. A medida de distância mais conhecida é a chamada distância

de Minkowski:

$$d_{L_p}(H_0, H_1) = \left[\sum_{i=1}^{n} |H_0(i) - H_1(i)|^p \right]^{1/p}. \quad (3.3)$$

Para $p = 2$, a Equação 3.3 leva à distância Euclidiana. Para $p = 1$, obtem-se a distância Manhattan. A métrica máxima é atingida para o limite da Equação 3.3, quando $p \to \infty$.

Este tipo de distância é sensível a deslocamentos no histograma, mesmo se os deslocamentos são pequenos. Isso vem do fato de que cada coluna deve casar exatamente com a sua correspondente. Quando se trata de imagens, este fato torna-se uma desvantagem considerável. Por exemplo, variações na iluminação deslocam o histograma de cor. No caso de histogramas de orientação — em análise de textura ou borda, por exemplo — problemas semelhantes persistem caso seja aplicado algum tipo de rotação às imagens. Então, a distância entre histogramas não é uma técnica robusta para lidar com imagens. Além disso, uma vez que essas distâncias são baseadas principalmente em diferenças absolutas, elas não são robustas a ruídos. Robustez a ruídos (especialmente do tipo impulso) pode ser alcançada com a chamada distância Jeffey, que por sua vez baseia-se na distância de Kullback-Leibler [Puzica et al., 1997], dada por:

$$d_J(H_0, H_1) = \sum_{i=1}^{n} \left[H_0(i) log(\frac{H_0(i)}{m_i}) + H_1(i) log(\frac{H_1(i)}{m_i}) \right], \quad (3.4)$$

onde $m_i = \frac{H_0(i)+H_1(i)}{2}$. Entretanto, esta distância ainda é sensível a deslocamentos das distribuições.

Uma outra medida de similaridade, robusta a ruído e a pequenas variações, é a chamada razão sinal/ruído, representada pela razão entre o "número de diferenças relativas" e o "número de variações" em uma distribuição:

$$d_{rsr}(H_0, H_1) = 10 log_{10} \left(\frac{Var[H_0]}{Var[H_0 - H_1]} \right). \quad (3.5)$$

Esta medida não é simétrica, o que significa que é sensível a pequenas perturbações nas localizações das distribuições. Entretanto, tal propriedade geralmente não é requerida em

sistemas RIBC. Também, essa medida é considerada como uma medida de similaridade global, uma vez que usa parâmetros globais dos histogramas (forma, distribuição, amplitude, etc.). Por outro lado, existe uma grande variedade de distâncias consideradas globais que são mais sensíveis a forma dos histogramas do que a seus valores exatos. Um exemplo é a "distância aberta" (do inglês *unfolded distance*), proposta por [Shen e Wong, 1983], e generalizada por [Werman et al., 1985], que consta primeiramente em ordenar os valores dos histogramas e em seguida calcular a soma das diferenças absolutas entre os pares de entradas correspondentes. Essa distância foi largamente testada para imagens de texturas em [Werman e Peleg, 1984]. Outro exemplo desta categoria de medidas é a distância EMD (do inglês, *Earth Mover's Distance*), idealizada pela primeira vez por [Rubner et al., 1998]. Essa medida é definida como sendo o custo mínimo para transformar um histograma em outro, e é uma variante do conhecido problema de transporte. Se H_0 e H_1 são os histogramas das imagens que se está querendo comparar, pode-se definir a distância entre eles de acordo com a Equação 3.6:

$$EMD(H_0, H_1) = \frac{custo\ mínimo}{min(área(H_0), área(H_1))}, \tag{3.6}$$

onde *custo mínimo* é o custo de transformar o histograma H_0 em H_1. Um algoritmo para calcular esse custo pode ser encontrado em [Russell, 1969].

3.7.2 O modelo vetorial

Uma medida de similaridade que considera a distância entre dois vetores é o chamado Modelo Vetorial. Este modelo utiliza um espaço v-dimensional, ou seja, $\vec{q} = (c_{1q}, c_{2q}, \ldots, c_{vq})$ e $\vec{p_j} = (c_{1j}, c_{2j}, \ldots, c_{vj})$, onde $\vec{q}$ é o vetor de características extraídas da imagem que a representa, e $\vec{p}$ é o vetor de características da imagem-alvo, no banco de dados, e c_{ij} é a característica associada a q e a p. Operações algébricas são efetuadas sobre tais vetores para obter uma ordenação dos documentos (imagens recuperadas) [Baesa-Yates e Ribeiro-Neto, 1999], [Salton e McGill, 1997], [Salton et al., 1997].

A similaridade entre a consulta e cada documento é calculada, neste trabalho, pela

correlação entre os vetores $\vec{q}$ e $\vec{p_j}$. Essa correlação, como foi visto na Seção 3.7.2 , é dada pelo coeficiente de correlação de Pearson, e representa o cosseno do ângulo formado entre os dois vetores, ou o produto interno entre eles, dado por:

$$sim(\vec{p}, \vec{q}) = \frac{p_j^t q}{|p_j||q|} = \frac{\sum_{i=1}^{n} c_{ij} \times c_i}{\sqrt{\sum_{i=1}^{v} c_{ij}^2} \times \sqrt{\sum_{i=1}^{v} c_{iq}^2}} \tag{3.7}$$

onde $|p_j|$ e $|q|$ são as normas (quadrática) do documento e da consulta. A norma da consulta não influencia o resultado da ordenação e a norma do documento permite a normalização no espaço de documentos.

As principais vantagens do modelo vetorial são: possibilidade de utilização de pesos para as características, melhorando o desempenho dos sistemas de RIBC; casamento parcial entre a consulta e representantes de imagens da coleção; e a ordenação das imagens na resposta pela similaridade parcial com a consulta. A principal desvantagem é considerar, como premissa, a independência das características. Esta premissa é simplificadora mas permite que os cálculos sejam funções lineares.

3.8 Conclusões

Foram apresentados diversos conceitos e definições que serão usados mais adiante. Objetos e cenas foram definidos dentro do contexto das bases de dados utilizadas; bem como documentos, coleção e consulta. Esses termos já são bastante conhecidos na área de recuperação de informação em geral, no entanto, para evitar ambiguidades, foram também definidos aqui.

Além dos conceitos de termos usados, também as idéias do que são uma rede bayesiana, e a decomposição em valores singulares, foram revisados brevemente. As redes bayesianas serão usadas mais tarde para modelar uma rede de crença para consulta de imagens; e a decomposição em valores singulares é uma técnica que divide o espaço — nesse caso, o vetorial — em três matrizes. Essas matrizes revelam relacionamentos entre os termos que podem ajudar na recuperação de informação ligadas a imagens.

Capítulo 4

Metodologias Propostas

Neste capítulo, são introduzidas e discutidas as soluções para os problemas propostos no Capítulo 1: o algoritmo *GRAS*, a ASL, e o modelo de redes bayesianas para ordenação dos documentos. O *GRAS* e a ASL são explicados à luz de exemplos não reais; a técnica de agrupamento é explicada através de seus algoritmos; e o modelo bayesiano à luz da dedução matemática de suas fórmulas. Os conceitos necessários para o entendimento destes assuntos estão contidos no Capítulo 3.

4.1 Algoritmo de extração de características espaciais: GRAS

4.1.1 Pré-processamento

O algoritmo *GRAS* funciona usando como entrada uma matriz $n \times n$ que contém as distâncias entre os nodos de um grafo direcionado, e fornece como saída um vetor de características desse grafo. Essa matriz é gerada a partir de uma imagem em tons de cinza. Logo, essas características também são características da imagem de entrada.

A seqüência de obtenção dessa matriz é a seguinte:

1. a imagem é segmentada em regiões;

2. o centróide de cada região é calculado;

3. as distâncias entre todos os centróides são calculadas;

O *GRAS* não precisa que as regiões da imagem que está sendo operada sejam previamente relacionadas ou fechadas por algoritmos especiais; o que ocorre em geral com diversos métodos de reconhecimento de objetos, ou recuperação de imagens. Isso constitui uma vantagem, uma vez que algoritmos de segmentação eficazes, ou de fechamento de segmentos e bordas, possuem custo computacional elevado. Os resultados mostram que, mesmo usando um pré-processamento simples, como o descrito a seguir, os resultados são satisfatórios.

Assim, para segmentar ma imagem em regiões, optou-se pela seguinte seqüência de três passos de operações sobre a imagem original [Gonzalez e Woods, 1992]:

1. Cálculo do mapa de bordas;

2. Eliminação de ruídos;

3. Absorção de pequenas regiões.

No Passo 1, é usado o filtro Sobel por ser uma operação simples de detecção de bordas. O problema que se tem ao usar uma operação como essa é que a imagem não ficará bem segmentada, pois ainda existirão ruídos, pequenas regiões isoladas e regiões que deveriam estar conectadas mas não estão. No entanto, é exatamente esse o intuito; pois assim, é reforçada a idéia de que o algoritmo funciona nessas condições.

No Passo 2, deseja-se excluir pequenas regiões isoladas, como ruídos. Então, usa-se um processo também simples de eliminação das mesmas, o que é feito da seguinte forma: sendo o mapa de bordas binário, todos os seus *pixels* possuem valores zero (0) ou um (1). Assim, os *pixels* cujos valores sejam iguais a 1, e que possuem na sua região 8-conectada mais de 5 *pixels* com valor igual a 0, são colocados para 0. O mesmo acontece se o valor de um *pixel* for 0 e possuir mais de 5 *pixels* com valor igual a 1. Nesse caso, o *pixel* é colocado para 1.

Finalmente, no Passo 3, uma certa quantidade de pequenas regiões deve ser absorvida por regiões maiores. Isso é conseguido através da operação morfológica "dilatação", com um elemento estruturante em cruz. Essa operação visa eliminar regiões geradas, não por ruídos, mas por elementos na imagem como pequenas sombras, reflexões da luz, etc. Essa operação também agrupa regiões vizinhas separadas por pequenos espaços.

A Fig. 4.1 mostra uma imagem de um objeto usada neste trabalho e a Fig. 4.2 ilustra o resultado das operações de pré-processamento discutidas acima.

Figura 4.1: Exemplo de uma imagem original de um objeto. Esse tipo de imagem possui fundo homogêneo e preto. As imagens foram capturadas em laboratório com iluminação controlada [Columbia, 2001].

A próxima etapa é calcular o centróide de cada região individual encontrada. Obviamente, caso a imagem ainda contenha algum *pixel* ou pequena região isolada (ruído, por exemplo), esta é considerada uma região válida como outra qualquer, influenciando "erroneamente" o resultado final. Entretanto, uma vez que o algoritmo tenta extrair características que são consideradas globais da imagem, essa influência pode ser minimizada.

As coordenadas cartesianas de cada centróide são consideradas as coordenadas cartesianas de cada nodo do grafo, e as suas distâncias são os valores da matriz que serve de

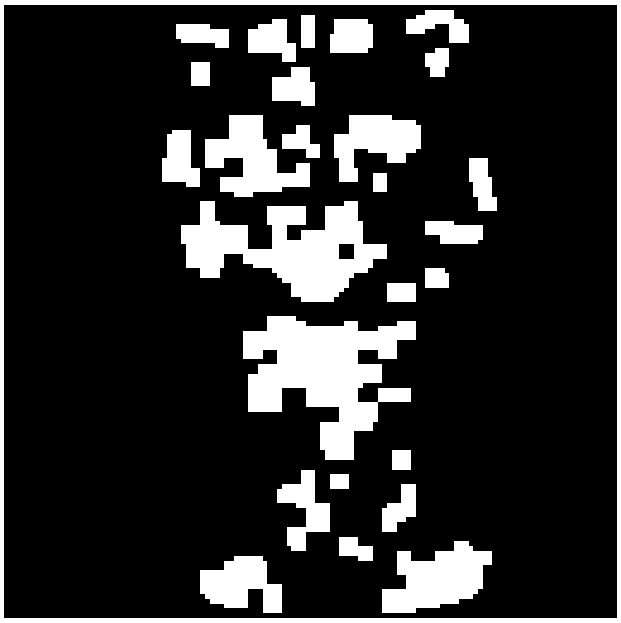


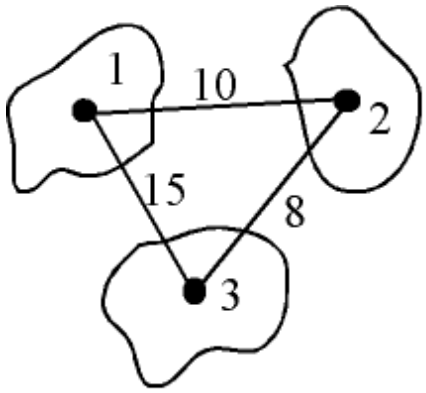
Figura 4.2: Imagem de regiões da Fig. 4.1, após o pré-processamento sugerido.

entrada para o algoritmo.

Para ilustrar o que foi dito, suponha as três regiões, rotuladas 1, 2 e 3, da Fig. 4.3a, com a posição e as distâncias dos centróides indicadas. A Fig. 4.3b exibe a matriz de distâncias, entrada para o algoritmo.

4.1.2 O Algoritmo *Graph Region Arrow Shot (GRAS)*

A idéia geral do algoritmo é ligar os centróides entre si, criando um grafo direcionado, conectando de cima para baixo, a partir do ponto mais alto, sempre conectando os pontos mais próximos até alcançar o mais baixo. O grafo construído possui uma série de características que podem ser extraídas durante a varredura. São características globais da estrutura geométrica do grafo — e conseqüentemente da imagem do objeto —, e que não são capturadas, geralmente, por algoritmos que baseiam-se apenas na identificação de bordas internas ou externas. Algumas características são invariantes ao ponto de partida do algoritmo (como NR, NC, EMG e TP); outras sofrem algumas variações (como NF e NS) — para essas siglas, ver Seção 4.1.3 —. No entanto, como são características globais,



(a)

	1	2	3
1	*	10	15
2	10	*	8
3	15	8	*

(b)

Figura 4.3: Exemplo de entrada para o *GRAS*: (a) 3 regiões (rotuladas com valores inteiros: 1, 2 e 3) após a segmentação. Os pontos pretos dentro de cada uma delas indicam os seus respectivos centróides. As distâncias entre elas também estão indicadas por valores inteiros (8, 10 e 15) nas arestas; (b) matriz de distâncias usada como entrada para o algoritmo, de acordo com a Fig 4.3(a). Nota-se que essa matriz é simétrica.

esse problema pode ser minimizado.

Estes tipos de características capturam informações do relacionamento global entre as regiões, e não da forma delas em si. Por esse motivo, o GRAS deve ser visto como um algoritmo que captura informações somente de relacionamento espacial, sendo portanto, melhor usado quando é combinado com outros métodos, como mostram os resultados experimentais no próximo capítulo.

Sejam p, q $(p < q)$, dois valores de rótulos de centróides quaisquer do conjunto de regiões que representa o objeto. Como o algoritmo baseia-se nas distâncias entre os nodos (centróides das regiões), é necessário definir um valor relativo para as distâncias. Esse valor, d, é tomado como sendo a média de todas as distâncias entre os nodos dividida por 2, que foi um valor tomado empiricamente que gerou os melhores resultados. O algoritmo funciona tendo como entrada os valores das distâncias entre os centróides e leva enconta o conjunto de regras a seguir:

1. se $p = 1$ e existe q a uma distância menor ou igual a d, aplica-se a Regra 2, caso contrário, aplica-se a Regra 3;

2. se um centróide $p \neq q$ está a uma distância menor ou igual a d de um centróide q, então cria-se uma aresta direcionada de p para q. Diz-se que p flecha q por limiar;

3. se um centróide p não possui nenhum outro centróide a uma distância menor ou igual a d, e q é o centróide mais próximo dele, então cria-se uma aresta direcionada de p para q. Diz-se que p flecha q por proximidade;

4. se um centróide $p \neq 1$ ainda não foi flechado, ele é flechado pelo centróide do conjunto de centróides já visitados mais próximo dele, em seguida aplica-se a Regra 2 ou 3 para p;

5. um ponto q não pode flechar um ponto p se p já tiver flechado q;

6. se p for o rótulo de maior valor, não flecha ninguém.

Para ilustrar o algoritmo, considere o conjunto fictício de regiões da Fig. 4.4a. Cada região foi rotulada com um valor inteiro de 1 a 15, da esquerda para a direita e de cima para baixo. Considere o conjunto de pontos da Fig. 4.4b. Cada um desses pontos representa o centróide de cada uma das regiões da Fig. 4.4a.

O algoritmo inicia a partir do centróide de coordenada mais alta. Nesse caso, 1. Para simular que um centróide q está a uma distância menor ou igual a d de um ponto p, na Fig. 4.4a os centróides foram ligados por pontilhados (- - - - -). Analogamente, se p e q estão separados por uma distância maior do que d, e q é o ponto maior do que p e mais próximo dele, na Fig. 4.4a, os centróides foram ligados por traço cheio (———).

Assim, no exemplo dado e de acordo com o conjunto de regras enumeradas acima, 1 flecha 2 e 4 por limiar (Regra 1, e depois 2). Em seguida, o algoritmo continua varrendo o conjunto de pontos a partir de 2. Os pontos que estão a uma distância menor ou igual a d de 2 são 1 e 5. Pela Regra 5, 2 não pode flechar 1, mas pela Regra 2 flecha 5 por limiar. Pela Regra 3, o centróide 3 não pode flechar 1 e 2, mas flecha 5 por proximidade. Continuando com o exemplo dado: 4, já flechado por 1, só pode flechar 7, que é o mais próximo (Regra 3); 5 flecha 6 por limiar (Regra 2); 6 flecha 8 por proximidade; 7 flecha 9,

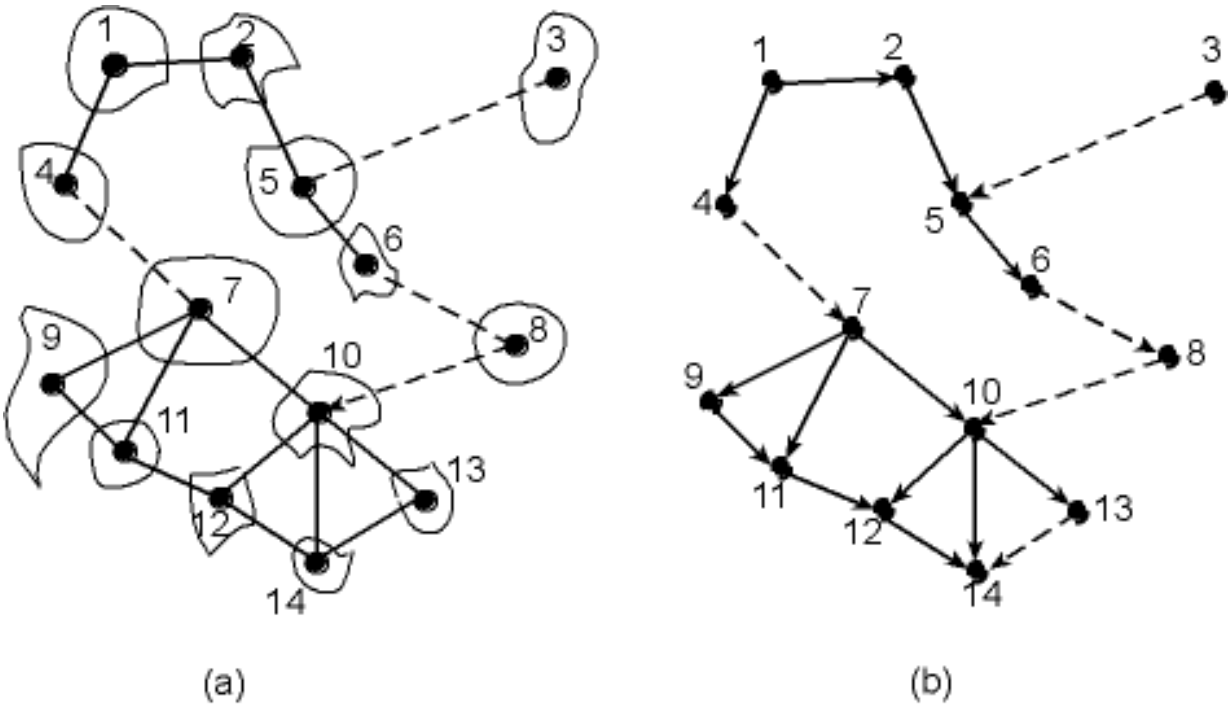


Figura 4.4: (a) um conjunto — fictício — de regiões de um objeto; (b) grafo conectado após o algoritmo *GRAS*.

10 e 11 por limiar; 8 flecha 10 por proximidade; 9 flecha 11 por limiar; 10 flecha 12, 13 e 14 por limiar; 11 flecha 12 por limiar; 12 flecha 14 por limiar; e finalmente, 14, pela Regra 6, não flecha ninguém.

Durante a varredura do algoritmo, várias características consideradas globais do objeto podem ser extraídas. No entanto, essas características não são únicas, e pode-se definir e extrair diversas outras. Apesar disso, para a finalidade do trabalho proposto aqui, elas são consideradas suficientes. A seguir, elas são definidas.

4.1.3 Características usadas

Nesta seção, são apresentadas as características extraídas pelo algoritmo proposto, durante a sua varredura. Foram escolhidas características que capturam o "comportamento" geométrico global do grafo — e conseqüentemente da imagem —.

No entanto, essas características não são únicas, e podem ser definidas e extraídas diver-

sas outras. Apesar disso, para a finalidade do trabalho proposto aqui, elas são consideradas suficientes.

- **Número de Regiões (NR)**: A maioria dos mais populares algoritmos de segmentação de imagens costuma gerar como saída, não exatamente uma única região que delimita o objeto em questão, separado de seu fundo, mas um conjunto de regiões que foram separadas umas das outras, Fig 4.2, mesmo pertencendo semanticamente a um mesmo objeto. Pode-se então tomar o número de regiões como sendo uma característica da imagem.

  NR é uma característica que dá a idéia de quão fragmentada pode ser a imagem. Imagens com grandes regiões homogêneas tendem a ter NRs baixos;

- **Número de Forks (NF)**: é o número de nodos em um grafo direcionado que possuem grau de saída pelo menos 2.

  Essa característica captura a quantidade de regiões que estão aglomeradas. Quanto maior o número dessas regiões, maior será o NF;

- **Número de Splashes (NS)**: é o número de nodos em um grafo direcionado que possuem grau de entrada pelo menos 2.

  O NS, como NF, também depende do número de regiões aglomeradas.

- **Número de Ciclos (NC)**: é o número de ciclos em um grafo, de acordo com a definição de [Cormem et al., 1997].

- **Eixo Médio Global (EMG)**: Cada região p da imagem possui um eixo maior, Em_i, e um menor, En_i, transversal a Em_i. Eixo médio global (EMG) é a razão entre a soma de todos os Em_i pela soma de todos os En_i.

  EMG dá uma idéia do quanto são alongadas as regiões da imagem.

- **Tempo de Preenchimento (TP)**: é o número de arestas de um grafo. Esse valor é proporcional à quantidade de regiões próximas umas das outras, de acordo com o limiar d.

	A	B	C	D	E	F	G	H	REL	CAS
Imagem1	x	x	x				x		R	
Imagem2				x^*	x			x^*		C
Imagem3			x	x^*				x^*	R	C

Tabela 4.1: Imagem-Consulta: I J H K D L

No caso do exemplo mostrado na seção anterior, os valores dessas características são: NR = 14; NF = 6 (5, 10, 11, 12, 14, 14); NS = 5 (1, 7, 7, 10, 10); NC = 5; TP = 18; EMG = $(E_{m1} + E_{m2} + ... + E_{m15})/(E_{n1} + E_{n2} + .. + E_{n15})$.

Essas características são componentes de um vetor que representa o objeto ou a imagem original. Para se definir a similaridade entre duas imagens de dois objetos, a métrica usada aqui é a do modelo vetorial, definida pela Equação 3.7.

4.2 Análise Semântica Latente (ASL)

4.2.1 Informações latentes e espaço latente

Quando um usuário faz uma busca em um sistema de RIBC, geralmente especifica imagens-consultas que não possuem todas as características que realmente poderiam ser necessárias para recuperar as imagens (da base de dados) que o atendam. Com o uso de modelos tradicionais[1], normalmente a associação entre uma imagem-consulta e uma imagem da base de dados é rigorosa: somente imagens que possuam estritamente características presentes na imagem-consulta serão recuperadas. Para ilustrar esse problema, é descrito o seguinte exemplo, com a ajuda da Tabela 4.1: Por simplificação, neste exemplo todas as imagens são representadas somente pelas características que estão contidas nelas. Para representar essas características, são usadas letras maiúsculas do alfabeto (A, B, C, etc.). Assim, a letra A, por exemplo, pode representar uma freqüência de uma determinada cor, um gradiente

[1]Neste trabalho, são chamandos de modelos tradicionais aqueles que usam puramente histogramas de cores, mapa de bordas ou informações de matriz de co-ocorrência, que fazem busca simplesmente minimizando a diferença entre vetores

de uma borda, uma direção, uma freqüência de direção, a entropia de uma matriz de co-ocorrência que representa textura, informações que dêem a idéia da distribuição espacial de regiões, etc.

Seja Im, a imagem-consulta fictícia utilizada para este exemplo, que é composta das seguintes características: "I J H K D L". Essa imagem deve ser confrontada com a base de dados de imagens. A Tabela 4.1 exibe uma base de dados que contém 3 imagens-alvo (Imagem1, Imagem2 e Imagem3). As letras A,B, C, D, E, F, G e H representam as características possíveis de ocorrer nessa base. Um x em uma célula indica que a característica da coluna correspondente ocorre na imagem da linha correspondente. Um R na coluna REL (RELevante) indica que o usuário julgou que a imagem correspondente é relevante para a sua consulta (nesse exemplo particular, as imagens 1 e 3 foram rotuladas como relevantes). Características que ocorrem na imagem-consulta e em uma imagem da base são indicadas por um x^* na célula correspondente; um C na coluna CAS indica que a imagem na linha correspondente "casou" com a imagem-consulta, e deve ser retornada para o usuário.

As imagens 1 e 2 ilustram classes de problemas com os quais o método proposto lida. A imagem 1 é uma imagem relevante, no entanto, não possui nenhuma das características que existem na imagem-consulta. Então, ela não será retornada para o usuário se forem usados métodos tradicionais. A imagem 2 não é relevante para o usuário, no entanto, contém características que ocorrem na imagem-consulta, e, desta forma, será retornada. Independente do fato de ela não ter sido julgada relevante, seu contexto exige que ela pertença ao conjunto de imagens que devem retornar. Então, fazer a interseção entre a imagem-consulta e a base de imagens pode não ser uma estratégia plausível devido ao fato de que a imagem 1 não será retornada.

Em suma, algumas imagens consideradas relevantes pelo usuário não são retornadas, uma vez que não possuem características existentes na imagem-consulta. Entretanto, ocorrem com relativa freqüência no contexto requerido.

As características que não pertencem à imagem-consulta, mas que fazem parte do contexto dela, são informações adicionais que normalmente não são levadas em conta quando o processo de busca é simplesmente fazer uma minimização entre a imagem-consulta e cada

imagem da base de dados. Essas informações adicionais são chamadas de "informações semânticas latentes" da imagem, e o espaço vetorial onde elas são encontradas é chamado de "espaço vetorial latente"; daí o seu nome: Análise Semântica Latente.

A inspiração deste trabalho para esse modelo veio da sua aplicação com relativo sucesso na área de busca em texto. Uma boa referência do uso de ASL para recuperação de documentos de texto pode ser encontrada em [Deewester et al., 1990].

4.2.2 O modelo de decomposição

Nesta seção, é descrita uma nova abordagem — e que é uma das propostas desta tese —- na área de RIBC, que serve como uma alternativa para aumentar a quantidade de informação — a informação latente — de uma imagem-consulta.

A idéia é encontrar um modelo que represente os relacionamentos entre características e imagens, de maneira que não somente características que apareçam na imagem-consulta sejam usadas para varrer a base de dados, mas também aquelas que ocorrem com relativa freqüência no contexto, as chamadas características latentes.

No caso de RIBC, o espaço vetorial da base de dados é usado para observar as ocorrências de características nas imagens. Considerando esse espaço, o objetivo aqui é usar um modelo matemático que explicitamente represente a estrutura semântica latente. A análise dessa estrutura é feita com o uso da decomposição em valores singulares (DVS), definida na Seção 3.5. Essa técnica está estritamente relacionada com um grande número de abordagens estatísticas em uma grande variedade de áreas, incluindo decomposição de autovalores, análise espectral e análise fatorial.

A decomposição, feita sobre o espaço vetorial de *imagens x características*, gera três matrizes especiais, através da DVS (ver Seção 3.5). Essas matrizes contém "valores singulares" e "vetores singulares", e "quebram" o relacionamento original entre imagens e características em uma série de componentes linearmente independentes chamados de fatores. Muitos desses componentes são muito pequenos, e podem ser ignorados, levando a um modelo aproximado em uma dimensão menor.

Nesse modelo reduzido, todas as similaridades *imagens x imagens* são agora representadas por valores dessa nova e menor dimensão. O resultado pode ainda ser representado geometricamente por uma configuração, na qual o produto interno ou o cosseno entre dois vetores representa a estimativa da medida de similaridade entre duas imagens.

Então, para os propósitos de RIBC, a DVS pode ser vista como uma técnica que gera um conjunto de valores de índices, chamados de fatores; onde cada imagem é representada por um conjunto desses fatores. Observe que, devido à redução da dimensão, imagens no espaço original com perfis de características diferentes, mas em contextos similares, são mapeadas para o mesmo vetor no espaço latente. Entretanto, esta é justamente a propriedade desejada; pois faz com que a DVS possa ser usada para modelar o problema da contextualização, discutida na Seção 1.2, uma vez que agrupa imagens por contextos de características similares (características latentes). Embora seus resultados estejam longe do ideal, a DVS é uma abordagem adequada para o problema, uma vez que permite a inclusão de informações que não foram necessariamente fornecidas diretamente pelo usuário. Além do que, o Capítulo 5 mostra que essa estratégia leva a uma melhoria tanto da precisão de busca (grau de certeza na recuperação) quanto da revocação (quantidade de imagens relevantes recuperadas).

De maneira superficial, os "fatores" gerados para o espaço latente podem ser pensados como "conceitos artificiais"; eles representam componentes comuns de muitas características e imagens. Cada imagem é então representada por um vetor de fatores que indicam seu relacionamento com outras imagens; isto é, o "significado" de uma imagem particular pode ser expressado por k fatores, ou equivalentemente, por uma localização de um vetor no espaço de fatores k-dimensional. Esta representação é econômica, no sentido de que os N índices originais são substituídos pelos $k \leq N$ melhores fatores. O objetivo aqui não é gerar um espaço onde esses fatores possam ser interpretados fisicamente, mas representar imagens e consultas de uma maneira que não seja a limitada, ambígua e redundante representação de características como descritores de imagens, simplesmente.

Nesta representação é possível reconstruir o espaço original *imagem x características*, a partir dos k-fatores, com relativa perda. É importante para o método que o espaço de

fatores k-dimensional não reconstrua o espaço original perfeitamente; isso porque o espaço original não é confiável. Primeiramente, deseja-se derivar um espaço que expresse o que é mais importante e influente no relacionamento entre características e imagens. Ao contrário de muitas maneiras de se usar análise fatorial, não se está necessariamente interessado aqui em reduzir a representação para uma dimensão menor — por exemplo, para 2 ou 3 fatores —, porque o objetivo não é ser capaz de visualizar ou entender topologicamente o espaço, mas sim, manipular informações latentes. Neste trabalho, acredita-se também que a representação do espaço conceitual, para qualquer coleção de imagens suficientemente grande, requererá mais do que a descrição de conceitos independentes, e o número de fatores ortogonais que serão necessários será proporcionalmente grande. Na realidade, relações conceituais entre características e imagens certamente envolvem estruturas mais complexas, incluindo, por exemplo, hierarquias locais e relações não lineares entre significados. Relações mais complexas podem ser geradas simplesmente aumentando a dimensão do espaço. Então, tem-se razão aqui em evitar tanto baixas quanto altas dimensões no espaço. No entanto, neste trabalho, os valores escolhidos são aqueles que dão os melhores resultados de busca (melhor precisão).

Como é possível então processar uma consulta nesta representação? Pode ser observado primeiro que cada imagem-consulta é representada por um vetor de fatores no espaço k-dimensional. Uma imagem-consulta inicialmente é um vetor de características. Para retornar um conjuto de imagens potencialmente relevantes, a imagem-consulta inicialmente é descrita nesse espaço, e os cossenos dos ângulos entre seu vetor de fatores e todos os outros vetores do espaço são calculados; aqueles com valor de cosseno alto são retornados como relevantes.

4.2.3 Detalhamento técnico

Esta seção detalha o aparato matemático por trás da DVS, utilizada para o modelo particular de estrutura latente. Como foi explicado na Seção 3.5, qualquer matriz retangular — por exemplo, uma matriz $i \times f$ de *imagens x características*—, X, pode ser decomposta

no produto de três matrizes

$$X = U_0 S_0 V_0^t,$$

de modo que U_0 e V_0 sejam ortogonais e S_0 seja diagonal. Essa operação é conhecida como a decomposição de X em seus valores singulares. A DVS é única para cada matriz X e, por convenção, os elementos da matriz diagonal são construídos de maneira que seus elementos sejam ordenados crescentemente, com o maior elemento ocupando a posição mais a esquerda superior e o menor a mais a direita inferior. A Figura 4.5 esquematiza esse processo.

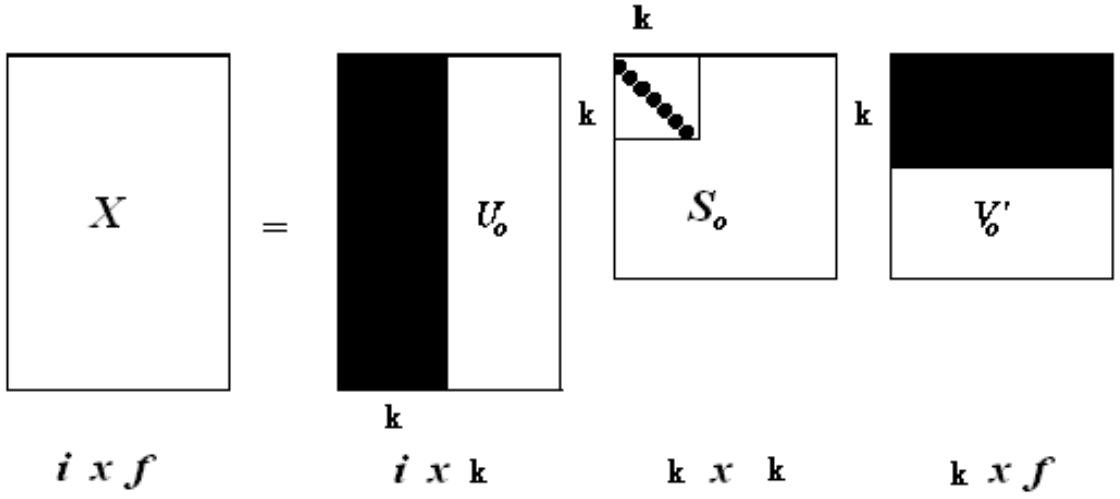


Figura 4.5: Decomposição do valor singular da matriz de *imagens x características*, X, onde: U_0 e V_0 são matrizes ortogonais, S_0 é a matriz de valores singulares, i é o número de linhas (imagens) de X, f o número de características, e k o número de fatores com os maiores valores.

Em geral, para $X = U_0 S_0 V_0^t$, as matrizes U_0, V_0 e S_0 devem todas ser de *rank* completo. A vantagem da DVS, entretanto, é que ela permite o uso de uma estratégia simples para representar um espaço aproximado do espaço original, com poucos e pequenos valores. Como os valores singulares de S_0 são ordenados por tamanho, os k valores maiores podem ser armazenados e os demais ignorados e colocados para zero. O produto das matrizes resultantes é uma matriz $\tilde{X}$, a qual é aproximadamente igual a X, mas com *rank* igual a

k. Uma vez que foram introduzidos zeros em S_0, a representação pode ser simplificada descartando-se as colunas e linhas iguais a zeros de S_0 com a obtenção de uma nova matriz diagonal, S, e descartando também as linhas e colunas correspondentes de U_0 e V_0, obtendo-se U e V, respectivamente. O resultado é um modelo reduzido:

$$X \approx \widetilde{X} = USV^t,$$

de *rank* igual a k. Este novo modelo (reduzido) é mostrado na Figura 4.6.

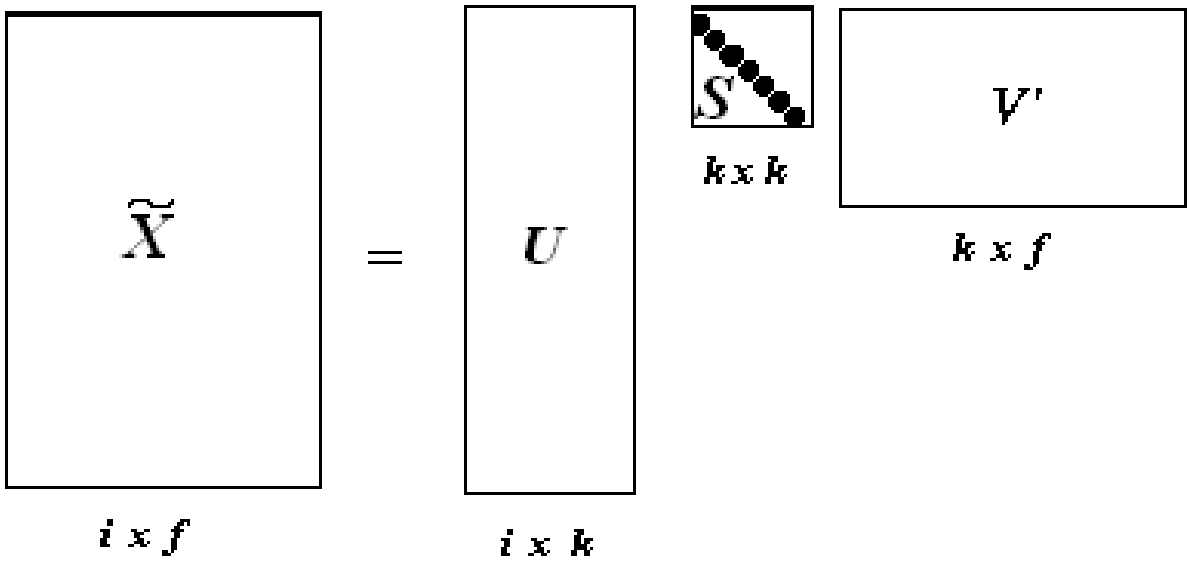


Figura 4.6: Decomposição do valor singular reduzido da matriz $\tilde{X}$. A notação pode ser encontrada na figura anterior.

Finalmente, de quanto a dimensão original é reduzida, isto é, a escolha de k, é um ponto crítico neste trabalho. Idealmente, deseja-se um valor de k grande o suficiente para representar a maior quantidade de informações possíveis, e pequeno suficiente de modo que não haja perda nem distorção de informações. A escolha do valor correto de k ainda é um problema em aberto. Na prática, usa-se atualmente o valor que leva aos melhores

resultados (precisão) possíveis de busca.

O produto interno entre as linhas da matriz $\tilde{X}$ reflete o quanto duas imagens possuem perfis de características similares. Então, a matriz $X\tilde{X}$ contém os produtos internos *imagem-imagem*.

4.2.4 Exemplo numérico do uso e efeito da ASL

Nesta seção, será dado um exemplo numérico com o uso da ASL em uma base vetorial formada pelo conjunto de 12 vetores de 9 dimensões cada. Analogamente com a base vetorial de imagens, cada um desses vetores pode ser visto como sendo um vetor de 9 características de cada imagem. Essa base, 12×9, de vetores é dada abaixo:

$$X = \begin{matrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{matrix}$$

De acordo com a Seção 3.5, qualquer matrix $i \times c$ pode ser decomposta em um conjunto de três matrizes:

$$X = T_0 S_0 D_0^t,$$

de modo que T_0 e D_0 possuem colunas ortonormais e S_0 é diagonal. Este processo é chamado de decomposição de X em valores singulares (DVS).

No exemplo dado, essas três matrizes geradas, considerando apenas duas casas decimais, pela DVS são as seguintes:

$$T_0 = \begin{matrix} 0.22 & -0.11 & 0.29 & -0.41 & -0.11 & -0.34 & 0.52 & -0.06 & -0.41 \\ 0.20 & -0.07 & 0.14 & -0.55 & 0.28 & 0.50 & -0.07 & -0.01 & -0.11 \\ 0.24 & 0.04 & -0.16 & -0.59 & -0.11 & -0.25 & -0.30 & 0.06 & 0.49 \\ 0.40 & 0.06 & -0.34 & 0.10 & 0.33 & 0.38 & 0.00 & 0.00 & 0.01 \\ 0.64 & -0.17 & 0.36 & 0.33 & -0.16 & -0.21 & -0.17 & 0.03 & 0.27 \\ 0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & -0.05 \\ 0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & 0.02 & -0.05 \\ 0.03 & -0.14 & 0.33 & 0.19 & 0.11 & 0.27 & 0.03 & -0.02 & -0.17 \\ 0.21 & 0.27 & -0.18 & -0.03 & -0.54 & 0.08 & -0.47 & -0.04 & -0.58 \\ 0.01 & 0.49 & 0.23 & 0.03 & 0.59 & -0.39 & -0.29 & 0.25 & -0.23 \\ 0.04 & 0.62 & 0.22 & 0.00 & -0.07 & 0.11 & 0.16 & 0.68 & 0.23 \\ 0.03 & 0.45 & 0.14 & -0.01 & -0.03 & 0.28 & 0.34 & 0.68 & 0.18 \end{matrix}$$

$$S_0 = \begin{matrix} 3.34 & & & & & & & & \\ & 2.54 & & & & & & & \\ & & 2.35 & & & & & & \\ & & & 1.64 & & & & & \\ & & & & 1.50 & & & & \\ & & & & & 1.31 & & & \\ & & & & & & 0.85 & & \\ & & & & & & & 0.56 & \\ & & & & & & & & 0.36 \end{matrix}$$

$$D_0 = \begin{matrix} 0.20 & -0.06 & 0.11 & -0.95 & 0.05 & -0.08 & 0.18 & -0.01 & -0.06 \\ 0.61 & -0.17 & -0.50 & -0.03 & -0.21 & -0.26 & -0.43 & -0.05 & -0.24 \\ 0.46 & -0.13 & 0.21 & 0.04 & 0.38 & 0.72 & -0.24 & 0.01 & 0.02 \\ 0.54 & -0.23 & 0.57 & 0.27 & -0.21 & -0.37 & 0.26 & -0.02 & -0.08 \\ 0.28 & 0.11 & -0.51 & 0.15 & 0.33 & 0.03 & 0.67 & -0.06 & -0.26 \\ 0.00 & 0.19 & 0.01 & 0.02 & 0.39 & -0.30 & -0.34 & 0.45 & -0.62 \\ 0.01 & 0.44 & 0.19 & 0.02 & 0.35 & -0.21 & -0.15 & -0.76 & 0.02 \\ 0.02 & 0.62 & 0.25 & 0.01 & 0.15 & 0.00 & 0.25 & 0.45 & 0.52 \\ 0.08 & 0.53 & 0.08 & -0.03 & -0.60 & 0.36 & 0.04 & -0.07 & -0.45 \end{matrix}$$

Considerando os erros por arredondamento, pode ser verificado facilmente que $X = T_0 S_0 D_0$, T_0 possui vetores-colunas ortogonais unitários, de modo que $T_0 T_0^t = I$, e da mesma forma $D_0 D_0^t = I$.

A seguir, X será aproximado usando apenas os dois principais valores singulares em S e os correspondentes vetores colunas de T e D. Neste modelo reduzido, $X \approx \tilde{X} = TSD^t$, onde

$$T = \begin{matrix} 0.22 & -0.11 \\ 0.20 & -0.07 \\ 0.24 & 0.04 \\ 0.40 & 0.06 \\ 0.64 & -0.17 \\ 0.27 & 0.11 \\ 0.27 & 0.11 \\ 0.30 & -0.14 \\ 0.21 & 0.27 \\ 0.01 & 0.49 \\ 0.04 & 0.62 \\ 0.03 & 0.45, \end{matrix}$$

$$S = \begin{matrix} 3.34 & \\ & 2.54 \end{matrix}$$

e

$$D^t = \begin{matrix} 0.20 & 0.61 & 0.46 & 0.54 & 0.28 & 0.00 & 0.02 & 0.02 & 0.08 \\ -0.06 & 0.17 & -0.13 & -0.23 & 0.11 & 0.19 & 0.44 & 0.62 & 0.53. \end{matrix}$$

Multiplicando as matrizes T, S e D^t, obtém-se finalmente:

$$\tilde{X} = \begin{matrix} 0.16 & 0.40 & 0.38 & 0.47 & 0.18 & -0.05 & -0.12 & -0.16 & -0.09 \\ 0.14 & 0.37 & 0.33 & 0.40 & 0.16 & -0.03 & -0.07 & -0.10 & -0.04 \\ 0.15 & 0.51 & 0.36 & 0.41 & 0.24 & 0.02 & 0.06 & 0.09 & 0.12 \\ 0.26 & 0.84 & 0.61 & 0.70 & 0.39 & 0.03 & 0.08 & 0.12 & 0.19 \\ 0.45 & 1.23 & 1.05 & 1.27 & 0.56 & -0.07 & -0.15 & -0.21 & -0.05 \\ 0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\ 0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\ 0.22 & 0.55 & 0.51 & 0.63 & 0.24 & -0.07 & -0.14 & -0.20 & -0.11 \\ 0.11 & 0.53 & 0.23 & 0.21 & 0.27 & 0.14 & 0.31 & 0.44 & 0.42 \\ -0.06 & 0.23 & -0.14 & -0.27 & 0.14 & 0.24 & 0.55 & 0.77 & 0.66 \\ -0.06 & 0.34 & -0.15 & -0.30 & 0.20 & 0.31 & 0.69 & 0.98 & 0.85 \\ -0.04 & 0.25 & -0.10 & -0.21 & 0.15 & 0.22 & 0.50 & 0.71 & 0.62 \end{matrix}$$

As buscas na base de dados, quando feitas sobre o espaço $\tilde{X}$ ao invés de X, levarão em conta todo o contexto da base, causando o efeito de uma busca por elementos mais freqüentes e não somente por características individuais, como ocorre quando a busca é feita em cima de X.

4.3 O Modelo de redes bayesianas

4.3.1 Modelo de rede de crença

O modelo de rede de crença para recuperação de informação, visual ou textual, segue a linha epistemológica na qual as probabilidades são interpretadas como graus de crença [Ribeiro-Neto e Muntz, 1996, Ribeiro-Neto et al., 2000]. Daí segue seu nome. A rede de crença é um grafo direcionado acíclico, no qual as probabilidades são interpretadas como graus de crença. Os nodos da rede representam variáveis aleatórias e os arcos representam o relacionamento entre elas, cuja intensidade é expressada pela probabilidade condicional da relação.

Uma simplificação comumente aceita nos modelos clássicos de recuperação de RIBC é que as imagens-consultas e imagens da base de dados são representadas por vetores de características e que estas são independentes entre si. Como resultado, imagens-consultas e imagens da coleção são modeladas como subconjuntos de características. Essa simplificação, claramente, pode levar a uma perda de informação semântica relevante.

No modelo proposto, cada característica possível de ocorrer em uma imagem, pode ser modelada como um nodo. Então, se C é o conjunto de todas as características possíveis, uma consulta Q é um subconjunto de C. Se cada imagem da base de dados for representada como um nodo da rede, de acordo com a Seção 3.6, um arco direcionado de cada característica em C para cada imagem da base deverá ser criado para representar a probabilidade de ocorrer uma imagem I_j qualquer, dado que ocorreu uma consulta Q. Esse raciocínio leva ao esquema da Fig. 4.7.

Como cada característica tem igual probabilidade de ocorrer em cada imagem da coleção, cada nodo da rede, do lado da imagem-consulta, está ligado a todos os nodos da coleção. No entanto, uma simplificação que normalmente é feita no modelo bayesiano é associar a cada arco do diagrama uma variável aleatória binária, para significar que a característica C_j associada ocorreu ou não — diz-se que está ativa ou inativa —. Então, quando uma consulta é feita, apenas os nodos referentes às características que ocorreram

na imagem-consulta estarão ativos.

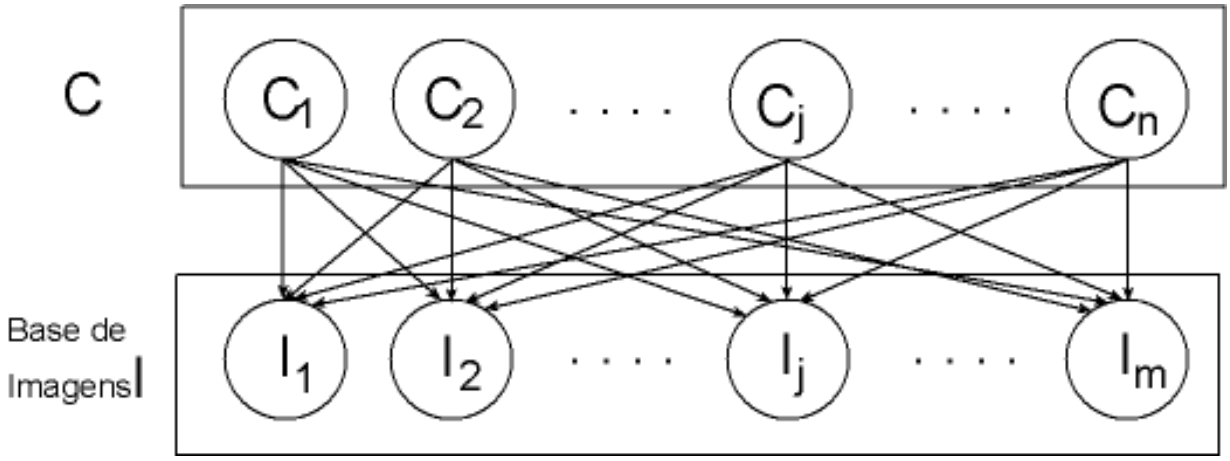


Figura 4.7: Rede bayesiana para RIBC. C é o conjunto de variáveis (características) que podem ocorrer; Cada I_j é uma imagem da coleção. Um arco de C_j para I_j representa a probabilidade de ocorrer a característica C_j (da imagem-consulta) na imagem I_j (da base).

A ordenação dos documentos em relação à consulta Q é dada por $P(I_j|Q)$. Pela definição de probabilidade condicional, a probabilidade $P(I_j|Q)$ pode ser calculada como segue [Russell e Norvig, 1995]:

$$P(I_j|Q) = \frac{P(I_j \wedge Q)}{P(Q)} \tag{4.1}$$

Como $P(Q)$ é uma constante para todas as imagens da coleção, assumindo que a probabilidade de ocorrer a consulta Q é sempre a mesma, independente da coleção, pode-se substituir $\frac{1}{P(Q)}$ por uma constante, ou seja $\frac{1}{P(Q)} = \eta$.

4.3.2 Rede de crença para o modelo vetorial

Os conceitos referentes ao modelo vetorial podem ser encontrados na Seção 3.7.2. A Equação (3.7) define como calcular a similaridade $sim(\vec{I_j}, \vec{Q})$ entre uma imagem I_j e uma consulta Q, segundo o modelo vetorial. No modelo de rede de crença, a similaridade entre a consulta Q e uma imagem da base I_j é dada pela probabilidade $P(I_j|Q)$, conforme a

Equação (4.1). Essa probabilidade $P(I_j|Q)$ pode ser feita equivalente a $sim(\vec{I_j}, \vec{Q})$, dada pelo modelo vetorial, através da seguinte especificação:

$$P(I_j|Q) = \begin{cases} sim(\vec{I_j}, \vec{Q}) & \text{se } |Q| - L \leq |I_j| \leq |Q| + L \\ 0 & \text{caso contrário} \end{cases} \tag{4.2}$$

onde L é um valor de limiar que estabelece que somente para os documentos cuja norma obedeça a $|Q| - L \leq |I_j| \leq |Q| + L$, a similaridade $sim(\vec{I_j}, \vec{Q})$ seja calculada. Isso diminui o número de imagens cuja similaridade é testada, diminuindo o tempo de processamento total. Esse limiar, L, é um valor empírico. Usou-se para os experimentos desta tese um valor que produzisse os melhores resultados possíveis. No entanto, algoritmos de otimização poderiam ser usados para encontrar automaticamente esse valor. Isso, no entanto, está fora do escopo desta tese.

Outro detalhe importante é com relação à faixa de valores possíveis (-1 a 1) para o modelo vetorial — e conseqüentemente para a semelhança entre duas imagens —, uma vez que eles representam o cosseno entre dois vetores. Para evitar valores negativos — que naturalmente não fazem sentido quando se trata de imagens —-, neste trabalho optou-se por normalizar os resultados para a faixa entre 0 e 1. Isso é equivalente à projeção para o primeiro quadrante de todos os valores de ângulos entre dois vetores que caírem fora da faixa de 0^0 a 90^0.

4.3.3 O modelo de rede bayesiana proposto para RIBC

A idéia aqui é combinar características tradicionais como histogramas de cor, histogramas de mapa de bordas, informações da matriz de co-ocorrência com Análise Semântica Latente, Agrupamento de imagens e informações espaciais globais (fornecidas pelo GRAS) em uma única rede bayesiana com o objetivo de se obter uma melhor ordenação (maior precisão e melhor revocação) das imagens recuperadas.

Cada conjunto de características da imagem (de forma, de cor e de textura) pode ser modelado em separado na rede proposta, em forma de vetor de características. Cada

uma das abordagens vistas nas seções anteriores são usadas de maneira combinada com técnicas tradicionais, para representar essas características. Assim, histogramas de cores, forma e textura podem ser combinados com informações espaciais, dadas pelo GRAS. Além do que, os espaços vetoriais gerados são decompostos com a ajuda da ASL e técnicas de agrupamento, como proposto anteriormente. O estudo da influência de todas essas abordagens no resultado final da busca é uma das contribuições desta tese.

Primeiramente, são criados em separado um vetor para características relacionadas à cor, outro para forma e outro para textura. As características de forma, a título de exemplo, são os gradientes do mapa de bordas das imagens; para as características de cor, o vetor contém somente o hitograma de cor quantizado no espaço HSV; e para características de textura são utilizadas informações da matriz de co-ocorrência das imagens em tons de cinza. A seguir, será descrita uma extensão do modelo de rede de crença [Ribeiro-Neto e Muntz, 1996] para RIBC. O modelo proposto, como foi dito, inclui 3 tipos de vetores de características. A Fig. 4.8 ilustra a proposta dessa rede.

Os termos da consulta são sempre as características de qualquer natureza, citadas acima. Cada um desses vetores é mapeado na rede pelos conjuntos **CF, CC e CT**, para representarem as características de forma, cor e textura, respectivamente. As variáveis IF_j, IC_j e IT_j representam as imagens da base de dados quando a busca for baseada na forma, cor e textura, respectivamente. Na verdade, as buscas são sempre sobre a mesma base de dados, mas por questão de visualização da Fig. 4.8, optou-se por repetir informação, apresentando os nodos referentes à base três vezes.

O diagrama da rede proposta é composto de quatro linhas e três colunas. A primeira linha, Q, representa a consulta; a segunda linha possui três vetores, CF, CC e CT, cada um representa os nodos da rede modelados para um conjunto de característica diferente: CF para características de forma, CC para característica de cor, e CT para característica de textura.

A terceira linha é composta também por três vetores, no entanto eles são os mesmos, representam as imagens da base de dados. Eles foram repetidos apenas por questão de entendimento do diagrama: se a linha três fosse composta apenas por um único vetor, muitos

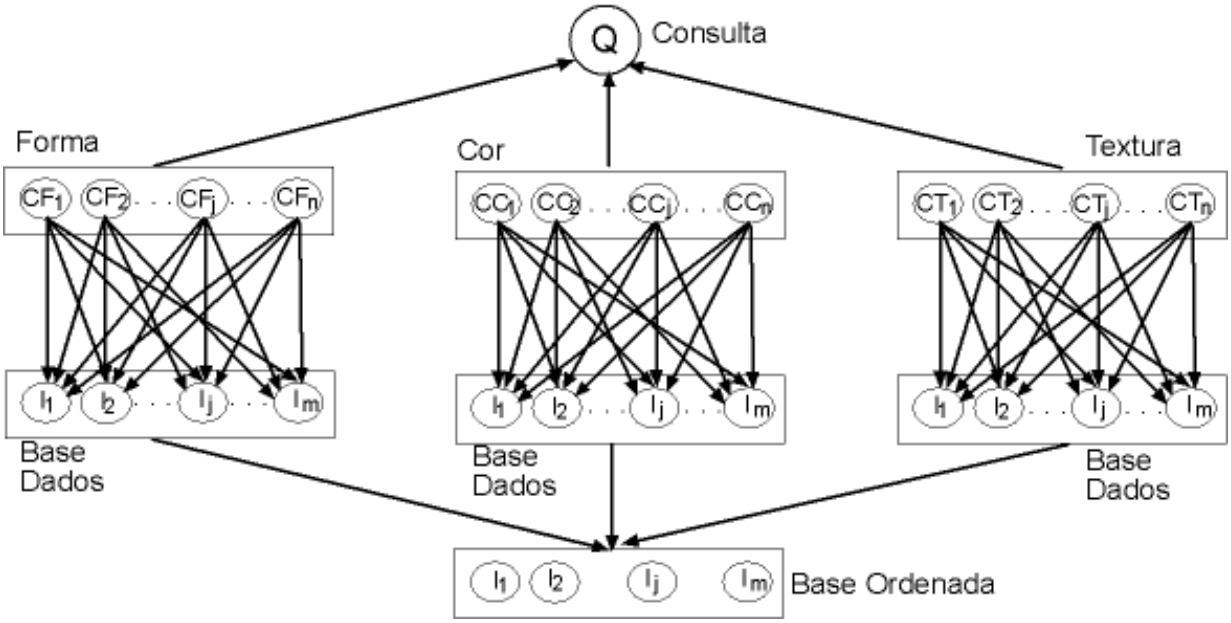


Figura 4.8: Rede bayesiana proposta para RIBC. Essa rede prossui três colunas e quatro linhas. A terceira linha, para evitar excesso de cruzamento de arcos (que dificultaria o entendimento do diagrama), tem informação redundante: a base foi repetida três vezes; uma para cada coluna que representa um tipo de busca distinto na rede (cor, forma e textura).

arcos partindo da segunda linha se cruzariam tornando a figura menos compreensível. Finalmente, a quarta linha é um único vetor. Ele representa o conjunto resposta que é a união disjunta dos conjuntos respostas de cada coluna, para cor, forma ou textura.

Os arcos partindo da segunda linha em direção à primeira, de acordo com a teoria de redes bayesianas vista na Seção 3.6, significam as probabilidades de ocorrer a consulta Q, dado que os conjuntos de características possíveis, CF, CC e CT ocorreram. Analogamente, cada arco que parte de um nodo C_j da segunda linha em direção a um nodo I_j na terceira linha significa a probabilidade de ser retornada a imagem I_j se a característica C_j ocorreu na consulta. Finalmente, os arcos saindo da terceira linha em direção à quarta significam as probabilidades de ocorrer a ordenação R, dado que retornaram os conjuntos de imagens produzidos separadamente por buscas baseadas na forma, cor e textura.

4.3.4 Equação geral para o cálculo da ordenação do conjunto resposta

Neste trabalho, considera-se que a ordenação das imagens do conjunto resposta é estabelecida pela probabilidade $P(I_j|Q)$, calculada pela Equação (4.2). No entanto, $P(I_j|Q)$ depende das diferentes características definidas neste trabalho, a saber: forma, cor e textura. Tais características são representadas na rede pelos conjuntos de termos **CF, CC e CT**, que geram as respectivas listas ordenadas de imagens **IF, IC e IT**. A idéia é desenvolver uma equação para 4.2 apenas com as informações que existem na rede. Assim:

$$P(I_j \mid Q) = \sum_{CF,CC,CT} P(I_j, CF, CC, CT|Q) = \sum_{CF,CC,CT} P(I_j|CF, CC, CT) \times P(CF, CC, CT|Q) \tag{4.3}$$

O segundo termo da Equação 4.3 pode ser ainda subdividido em:

$$P(CF, CC, CT|Q) = P(Q|CF, CC, CT) \times \frac{P(CF, CC, CT)}{P(Q)} \tag{4.4}$$

O termo fracionário da Equação 4.4 pode ser representado por uma constante, por exemplo, η, e o outro termo multiplicativo, por simplificação, pode ser dado da seguinte maneira:

$$P(Q|CF, CC, CT) = \begin{cases} 1 & \text{se todos os termos da consulta forem ativos}^2 \\ 0 & \text{caso contrário} \end{cases} \tag{4.5}$$

Assim, o único termo que resta para ser calculado é o primeiro termo da Equação 4.3, que é modelado como uma disjunção das crenças geradas por IF, IC e IT. Assim, pode-se escrever a Equação 4.3 como segue:

$$P(I_j \mid Q) = \eta \left[1 - P(\overline{IF}_j \mid CF) \times P(\overline{CC}_j \mid CC) \times P(\overline{CT}_j \mid CT)\right] \tag{4.6}$$

Cada uma das parcelas da Equação (4.6) pode ser representada em separado, por exemplo, $P(\overline{IF}_j \mid CF)$ pode ser representada por:

$$P(\overline{IF}_j \mid CF) = \frac{P(\overline{IF}_j \wedge CF)}{P(CF)} \tag{4.7}$$

Como $\frac{1}{P(CF)} = \eta$ é constante para todas as imagens da coleção, a Equação (4.7) pode ser representada como segue:

$$P(\overline{IF}_j \mid CF) \;\; = \;\; \eta \, P(\overline{IF}_j \wedge CF)$$

Esse mesmo raciocínio vale para as contribuções das outras parcelas da Equação (4.6), que pode então ser reescrita como segue:

$$P(I_j \mid Q) = \eta \, [1 - P(\overline{IF}_j \mid CF) \times P(\overline{IC}_j \mid CC) \times P(\overline{IT}_j \mid CT)] \tag{4.8}$$

$$P(I_j \mid Q) = \eta \, [1 - (1 - P(IF_j \mid CF)) \times (1 - P(IC_j \mid CC)) \times (1 - P(IT_j \mid CT))] \tag{4.9}$$

Note que o modelo bayesiano proposto aqui permite considerar cada consulta em separado. Por exemplo, para considerar somente os termos da consulta referentes às características de forma, basta definir na Equação (4.9): $P(IC_j \mid CC) = 0$ e $P(IT_j \mid CT) = 0$.

Adicionalmente, combinações de características uma a uma, duas a duas, também podem ser consideradas seguindo raciocínio análogo.

Capítulo 5

Experimentos

5.1 Justificativa da condução dos experimentos

Um sistema de recuperação de informação visual, ao receber como entrada a imagem de um objeto, pode ser capaz de responder a dois tipos de questões específicas: (a) semanticamente falando, que objeto a imagem representa ? e (b) quais as imagens que mais se assemelham à entrada? Essas duas questões são fundamentais e podem ser aplicadas a sistemas com objetivos diversos. O primeiro tipo de questão, (a), pode ser necessária quando se deseja saber precisamente a natureza semântica de um objeto, como por exemplo, que objeto aparece em determinada cena. A questão tipo (b) está mais ligada à área de recuperação de informação visual, como um sistema RIBC, ênfase desta proposta. Pode-se, por exemplo, querer saber, não exatamente a natureza do objeto de entrada, como no caso específico da questão (a), mas quais (ou quantas) imagens representam objetos que mais se assemelham à imagem do objeto dada como entrada. Essas duas questões são representativas do universo de duas áreas bem distintas: a questão tipo (a) é típica de Inteligência Artificial, enquanto que a questão tipo (b), por sua vez, é típica de RIBC. Analisando cada uma dessas questões, nota-se claramente a diferença entre "reconhecer" a imagem de um objeto apresentado ou "recuperar" as imagens que mais se assemelham a uma dada consulta. A diferença entre essas duas questões é fundamental, uma vez que a tarefa de reconhecimento inclui um

processo cognitivo, e a tarefa de "recuperação" consta apenas de buscar os similares, não necessariamente importando saber as suas naturezas semânticas. Entretanto, embora os métodos propostos aqui tenham sido projetados para RIBC, podem ser utilizados, com certas restrições, em aplicações que necessitem de reconhecimento.

À luz dessas idéias, os experimentos foram conduzidos com o uso de duas bases de dados distintas; uma com o intuito de medir as propostas com relação a sistemas puramente de recuperação de imagens semelhantes, outra com o intuito de medir os sistemas do ponto de vista de reconhecimento de padrões. Todos os resultados são exibidos com o uso dos gráficos de "Precisão x Revocação". Finalmente, várias classes de experimentos foram conduzidas combinando-se todas ou parte das propostas apresentadas.

5.2 Bases de dados utilizadas

Como foi falado anteriormente, todos os experimentos foram conduzidos com duas bases de dados distintas. A primeira delas, é a base de dados da Columbia [Columbia, 2001]. Esta base contém 100 classes de imagens de objetos. Cada classe possui 72 vistas do mesmo objeto. Cada vista é rotacionada de 5 graus da vista anterior, iniciando-se em 0 e finalizando em 355, perfazendo um total de $100 \times 72 = 7200$ imagens coloridas. Convencionou-se, neste trabalho, chamar esta base de "base controlada", uma vez que as imagens foram capturadas em laboratório com parâmetros como intensidade de luz, reflectância e ângulo de câmeras controlados. Um exemplo de classes e vistas contidos nesta base pode ser observado na Fig. 5.1.

A segunda é uma base de dados de imagens naturais, e foi construída na Universidade de ENSEA [Fournier et al., 2001], contendo 14 classes semanticamente distintas: aviões, carros, elefantes, vistas aéreas, ursos, leões, pessoas, retratos, peixes, répteis, paisagens, texturas, pôr-do-sol e setas, num total de 1200 imagens de cenas naturais coloridas. Uma vez que essa classe não possui imagens que podem ser consideradas vistas de um mesmo objeto, como na base Columbia, cada classe possui um número variado de membros. Um exemplo de classes e membros contidos nesta base podem ser observados na Fig. 5.2.

Figura 5.1: Base de dados controlada da Columbia. Aqui estão 8 das 100 diferentes classes de imagens de objetos (mostradas na horizontal), cada uma com 4 das 72 vistas possíveis (mostradas na vertical).

Assim, as propostas apresentadas nesta tese foram testadas em um total de 8400 imagens, estando de acordo com a literatura.

Com o uso dessas duas bases, ratifica-se que o intuito é testar as metodologias sob dois paradigmas distintos, mas que no entanto andam juntos : "reconhecimento" de objetos e recuperação de imagens semelhantes.

5.3 *Hardware* Utilizado

Todos os métodos estudados aqui, bem como os experimentos conduzidos, foram implementados em um computador bi-processado com processadores Intel(R)XEO com 2 GHz de *clock* e 1 GB de Memória RAM, com disco rígido de 30 GB e sistema operacional *Windows 2000*. O ambiente de desenvolvimento foi o Matlab 5.5 com os pacotes de processamento de images, estatística e processamento de sinais incluídos. Nesta máquina, e com esta configuração, o tempo de processameto médio *off-line* de toda a base de dados é em torno de 2 horas para a base da Columbia e 20 minutos para a base natural. O processamento *online* de todas as consultas gira em torno do mesmo tempo para ambas as bases. Uma única consulta, dependendo do método utilizado, fica em torno de 1 a 2 segundos para a base natural e cerca de 2 a 5 segundos para a base da Columbia.

5.4 O gráfico precisão x revocação

Para avaliar quantitativamente a eficiência dos métodos propostos, todos os resultados são mostrados com a ajuda dos gráficos clássicos de precisão e revocação [Smith, 1998]. Se **A** é um conjunto de imagens relevantes contido na base de dados, e **B** é o conjunto de imagens recuperadas dessa base, a precisão e a revocação são definidas da seguinte maneira:

- precisão $= \frac{|A \cap B|}{|B|} = \frac{\text{número de imagens relevantes recuperadas}}{\text{número de imagens recuperadas}}$
- revocação $= \frac{|A \cap B|}{|A|} = \frac{\text{número de imagens relevantes recuperadas}}{\text{número de imagens relevantes}}$

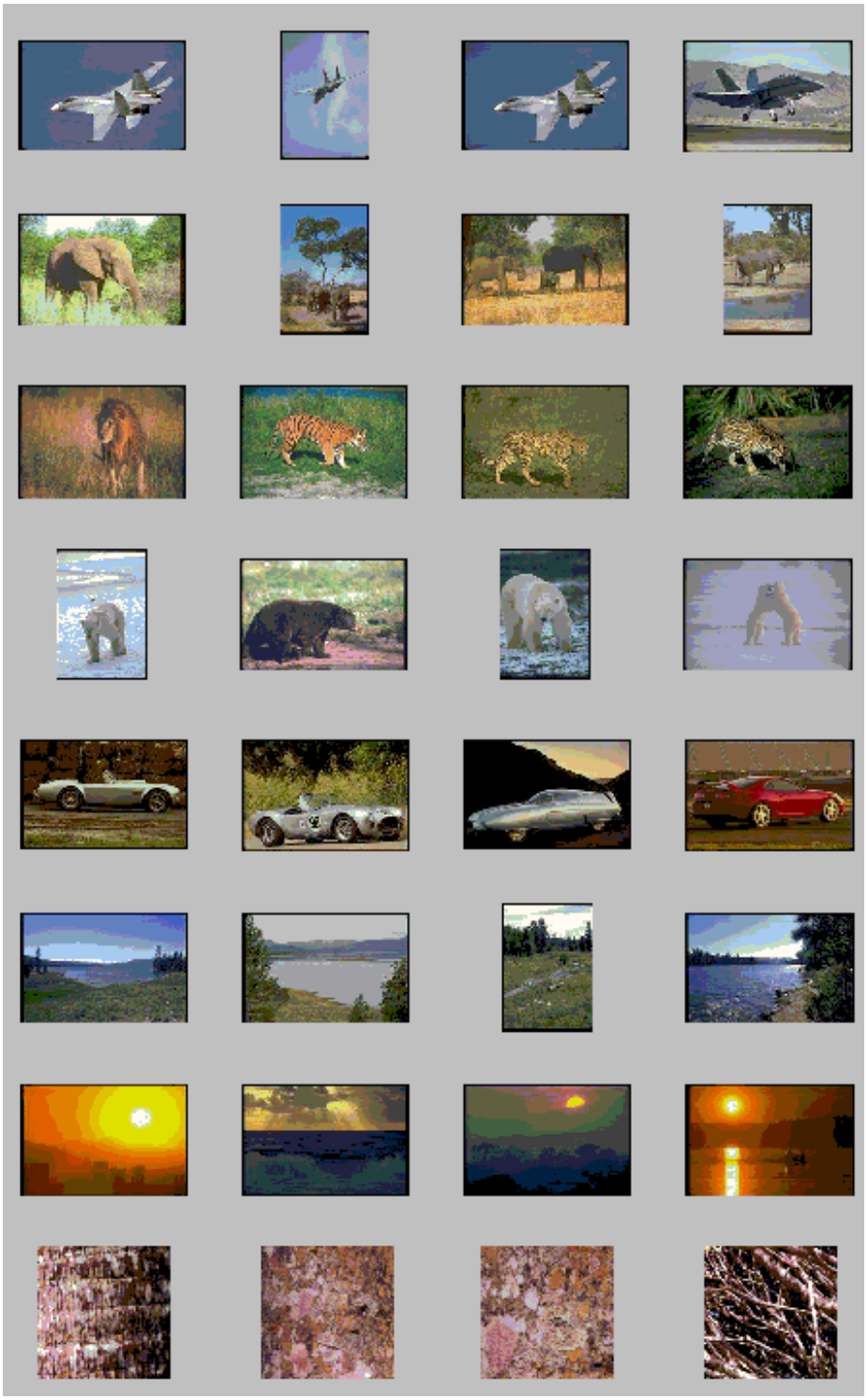


Figura 5.2: Base de dados natural. Aqui são mostradas 8 das 14 diferentes classes de imagens de cenas naturais (mostradas na horizontal). De cima para baixo, cada linha apresenta 4 imagens de uma classe distinta, na seguinte ordem: aviões, elefantes, felinos, ursos, carros, paisagens, pores-do-sol e texturas.

Ordem	Imagem	Relevante	Ordem	Imagem	Relevante
1	c_{11}	R	9	c_{24}	
2	c_{13}	R	10	c_{15}	R
3	c_{22}		11	c_{33}	
4	c_{12}	R	12	c_{35}	
5	c_{21}		13	c_{34}	
6	c_{14}	R	14	c_{23}	
7	c_{32}		15	c_{25}	
8	c_{31}				

Tabela 5.1: Exemplo de um resultado de uma consulta cuja imagem exemplo é c_{1j}. As cinco imagens relevantes da classe c_{1j} ocorrem nas dez primeiras posições.

A definição do que é relevante ou não, é uma tarefa difícil que depende da subjetividade humana. Por exemplo, na base de dados Columbia uma imagem é relevante se ela pertence à categoria de um conjunto de vistas para um mesmo objeto. Entretanto, para a base de dados natural, a tarefa de definir se uma determinada imagem é ou não uma boa resposta para uma determinada consulta é muito mais complicada. Assim, essa base foi manualmente classificada em 14 categorias distintas — oito das quais são mostradas na Fig. 5.2 — (aviões, felinos, ursos, carros, setas, etc.).

Exemplo do cálculo da precisão e revocação

Suponhamos uma determinada base de dados que possua apenas 3 classes de imagens, cada uma com 5 membros distintos. Para facilitar a explicação, será usada uma notação c_{ij} para significar que c é o membro j da classe i. Então $i = 1, .., 3$ e $j = 1, .., 5$. Supondo uma consulta da classe c_{1j} que retornou a ordenação da base de dados, mostrada na Tabela 5.1.

Nessa tabela, um R na coluna Relevante significa que a imagem c_{ij} da linha correspondente é relevante para a consulta[1]. Assim, para uma revocação de 1/5 (posição 1 da tabela) a precisão é $1/1 = 100\%$, uma vez que a primeira imagem relevante ocupa a posição 1 de 1.

[1]Uma vez que o índice i de c_{ij} já indica se essa imagem é relevante ou não, a coluna Relevante é desnecessária. No entanto, ela é mantida por questões de compreensão do exemplo.

Para uma revocação de 2/5, a precisão será de 2/2 = 100%, significando que as duas primeiras imagens relevantes ocupam as duas primeiras posições nessa ordenação. Então, se as cinco imagens relevantes ocorressem nas cinco primeiras posições, teria-se 100% de precisão para todas as revocações (1/5, 2/5, 3/5, 4/5 e 5/5).

No entanto, para uma revocação de 3/5, a próxima imagem relevante ocorre somente na quarta posição. Portanto, para essa revocação, 3/5, a precisão é de 3/4, significando que entre as cinco primeiras imagens ordenadas, apenas 4 são relevantes.

Seguindo esse raciocínio, a precisão para as revocações de 4/5 e 5/5 são 4/6 e 5/10, respectivamente.

5.5 Metodologia para avaliação

A metodologia para avaliação adotada é simples: cada imagem das bases de dados é imagem-consulta uma vez, perfazendo-se assim 8400 consultas. Para cada imagem-consulta de uma das bases (natural ou controlada), a ordenação de toda a base é retornada como resposta. No final, a média aritmética simples por classes é calculada e a Precisão x Revocação (PR) são usadas como medidas quantitativas.

Foram tomados tantos gráficos PR quantas foram as combinações dos experimentos. Por exemplo, um gráfico PR foi construído para exibir o desempenho da influência da ASL quando histogramas de cores são usados; o mesmo é feito quando utiliza-se ASL com histogramas do mapa de bordas, histogramas de informações de textura, e o algoritmo GRAS, etc. Finalmente, gráficos PRs são calculados para as mesmas condições, só que com o uso de redes bayesianas.

5.6 Métodos clássicos e novos métodos implementados

Por questões de simplicidade, as 14 classes da base natural serão chamadas pelos nomes: aviões, carros, elefantes, vistas aéreas, ursos, leões, pessoas, retratos, peixes, répteis, paisagens, texturas, pôr-do-sol e setas. Em contrapartida, devido a base Columbia ser relativamente grande (com 72 classes) e seus objetos possuirem nomes ambíguos, as suas classes serão denominadas por números inteiros.

5.6.1 HSV-162 e HSV-162 + GRAS

Primeiramente, será mostrada uma comparação do desempenho da estratégia de histograma de cores contra essa mesma estratégia combinada com características extraídas pelo algoritmo GRAS. A abordagem clássica que utiliza histogramas de cores é descrita por [Smith e Chang, 1996b]. Ela consta primeiramente em transformar a imagem para o espaço HSV (*Hue, Saturation, Value*) com 18 tons para o H, 3 para o S, 3 para o V e 4 para tons de cinza, gerando 18 x 3 x 3 + 4 = 166 possíveis valores para cada *pixel* em 18 + 3 + 3 + 4 = 28 entradas para o histograma. Por esse motivo, essa abordagem é conhecida como histograma-166. No presente trabalho, foi implementada uma variação dessa estratégia, que utiliza H = 18, S = 3 e V = 3. Os tons de cinza foram ignorados uma vez que aqui as imagens monocromáticas não são consideradas. Assim, foi implementado um histograma com 18 x 3 x 3 = 162 possíveis valores de *pixels*.

Por outro lado, o algoritmo GRAS foi usado para encontrar as características de regiões do grafo gerado a partir das imagens. Essas características são exatamente as 6 descritas na Seção 4.1.3: NR, NF, NS, NC, EMG e TP.

Um vetor de característica para uma imagem é então composto com as 18 + 3 + 3 = 24 entradas do histograma-162 mais as 6 características do GRAS, gerando um vetor de 24 + 6 = 30 entradas. .

Para efeito de comparação, as abordagens usando histograma-162 e histograma-162

Revocação	HSV-162	HSV-162 + GRAS	Ganho (%)
0.10	0.9912	0.9912	0
0.20	0.8370	0.8384	0.1654
0.30	0.7621	0.7636	0.1983
0.40	0.7291	0.7298	0.1068
0.50	0.6788	0.6805	0.2421
0.60	0.5523	0.5550	0.4852
0.70	0.3328	0.3341	0.3612
0.80	0.1991	0.1997	0.3042
0.90	0.0606	0.0607	0.0462
1.00	0.0277	0.0277	0.1053
Média	0.5171	0.5181	0.1909

Tabela 5.2: Desempenho do histograma-162 e do histograma-162 combinado com o GRAS para 72 consultas da classe 1. Embora as revocações tenham sido tomadas nos pontos 1/72 ... 72/72 (72 pontos), a tabela exibe uma interpolação linear para os dez pontos de revocação mostrados.

combinado com GRAS foram executadas em ambas as bases de dados, Columbia e natural. Primeiramente, uma única tabela PR para o histograma-162 foi gerada a partir da média de 72 consultas (uma consulta para cada imagem da classe 1). Assim, cada uma das 72 imagens da classe 1 é tomada como consulta 1 vez e confrontada com as 7200 imagens da base. O mesmo tipo de tabela, PR, foi gerada para o histograma-162 + GRAS. As duas tabelas foram combinadas em uma só, Tabela 5.2, onde também são mostrados os ganhos por revocação e ganho final médio. A primeira observação que pode ser tirada da Tabela 5.2 é o pequeno ganho em todos os pontos de revocação. Embora seja pouco, o ganho final médio, 0.19%, não representa uma média de perdas e ganhos, uma vez que o histograma-162 + GRAS superou o histograma-162 em todos os pontos de revocação, o que pode indicar, dentre uma base de dados de 7200 imagens, um ganho significativo para o usuário. Sem dúvida, esse pequeno ganho é devido ao aumento da quantidade de informação fornecida pelo GRAS. É, portanto, uma indicação prática de que a combinação de informações de forma com informações de cores pode levar a melhores resultados.

Outro ponto que é possível observar é que a quantidade de informação de cores usada é muito superior a pequena quantidade informação de forma. Isso explica o pequeno ganho,

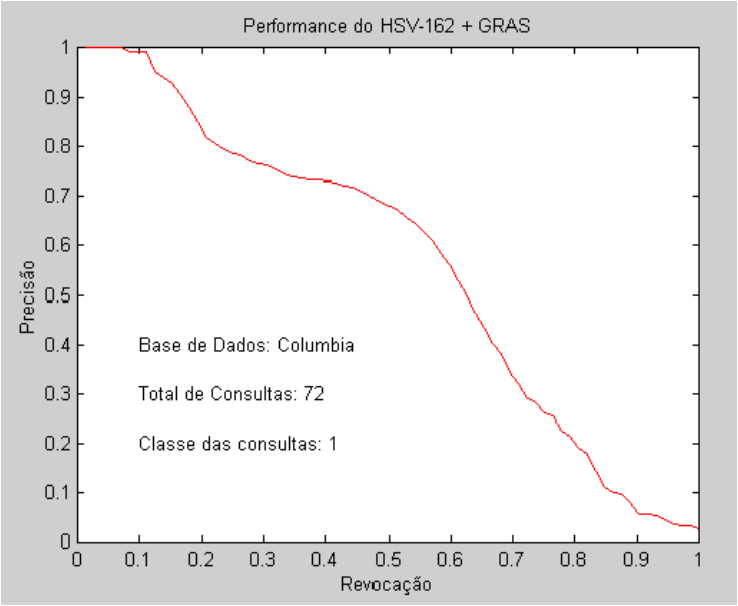


Figura 5.3: Desempenho do histograma-162 combinado com o GRAS para a classe 1 da base de dados Columbia. Cada uma das 72 vistas da classe foi tomada como consulta confrontada com as 7200 imagens da base. O gráfico é a média dessas 72 consultas.

no entanto, aumentar a quantidade de informação de forma e diminuir a de cor é uma estratégia que precisa ser investigada futuramente.

A queda "constante" da evolução das precisões, exibida no gráfico da Fig. 5.3, em relação às revocações é outro fator importante a ser considerado. Isso pode ser um reflexo do fato de ser uma base controlada, onde as vistas variam em taxas quase constantes de parâmetros (ângulo de visão, luminosidade, etc.). Entre as revocações 0.2 e 0.7, o gráfico exibe uma subita melhora no desempenho, sugerindo que as consultas nesta faixa eram imagens com melhor conjunto de relevantes, isso é comprovado observando como as vistas foram fotografadas; como há um giro de 360º, é natural que uma faixa de imagens semelhantes se repitam, gerando um desempenho periódico. No entanto, o desempenho sempre cai, devido ao aumento da revocação. O mesmo foi feito para a base natural; foram feitos experimentos para cada uma das 14 classes da base. Como na base da Columbia, cada membro de cada classe foi tomado como consulta e calculada a média global por classe. A Tabela 5.3 mostra o resultado médio de 175 consultas da classe carro da base natural. Ao contrário da classe 1 da base Columbia, claramente a combinação das características

Revocação	HSV-162	HSV-162 + GRAS	Ganho (%)
0.10	0.3421	0.6192	44.7513
0.20	0.3037	0.5307	42.7737
0.30	0.2740	0.4093	33.0687
0.40	0.2476	0.3202	22.6733
0.50	0.2286	0.2662	14.1274
0.60	0.2114	0.2392	11.6221
0.70	0.1982	0.2240	11.4758
0.80	0.1876	0.2046	8.3089
0.90	0.1779	0.1882	5.4729
1.00	0.1578	0.1604	1.6209
Média	0.2329	0.3162	26.3465

Tabela 5.3: Desempenho do histograma-162 e do histograma-162 combinado com o GRAS para 175 consultas da classe carro da base natural. Como na tabela anterior, embora as revocações tenham sido tomadas nos pontos 1/175 ... 175/175 (175 pontos), a tabela exibe uma interpolação linear para os dez pontos de revocação mostrados.

extraídas pelo GRAS com o histograma-162 superam o desempenho no caso de ser usado somente o histograma-162. Em alguns pontos de revocação o ganho chega a ser de quase 50%. O ganho médio, em torno de 27%, é devido às características da base natural. Ela possui muitas informações globais que geralmente não são capturadas somente com o uso de histogramas.

O gráfico da Fig. 5.4 compara o desempenho do histograma-162, quando ele é usado sozinho e quando é combinado com o GRAS.

O mesmo foi feito para as demais classes. Na Tabela 5.4 são mostradas as precisões e revocações para a classe texturas, podendo serem vistos ganhos de até 70%, com um ganho médio de 55%. Sendo essa classe, portanto, a mais adequada para o uso do GRAS. A Figura 5.5 mostra o comportamento global para todos os pontos. Com cerca de 70% de revocação, o desempenho cai muito, praticamente se igualando ao desempenho do histograma-162. Esse fato pode indicar o seguinte. Para grandes revocações, muitas imagens da base de dados se assemelham muito às imagens de pôr-do-sol, dificultando os resultados. Este fato relembra o problema da classificação manual da base; muitas imagens que possuem um pôr-do-sol e um carro poderiam ser classificadas tanto em uma classe quanto em outra.

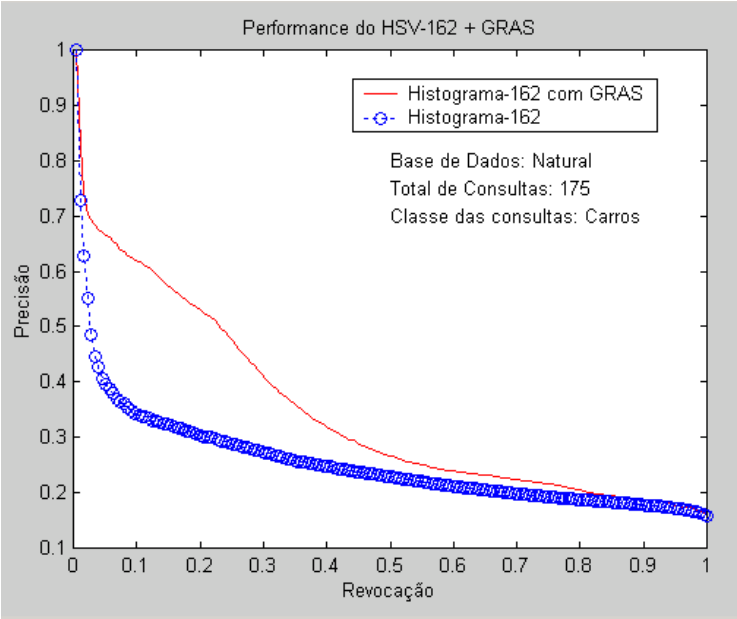


Figura 5.4: Desempenho do histograma-162 combinado com o GRAS para a classe carro da base natural.

Como exemplo, ver Fig. 1.2 na Seção 1.1.1.

A Tabela 5.5 apresenta os mesmos experimentos para a classe pôr-do-sol da base natural. Outra vez, a combinação de informações espaciais globais, fornecidas pelo GRAS, com histograma-162 supera em praticamente todos os pontos de revocação o histograma de cores. Apenas para uma revocação de 100% há uma perda em torno de 3% da técnica combinada em relação à tecnica clássica. Essa perda pode vir mais uma vez do fato de a base ter sido classificada manualmente, gerando possíveis imagens que podem ter sido classificadas como Por-do-Sol, no entanto, seus atributos espaciais classificam-na como pertencente a uma outra classe. Novamente, a Fig. 5.6 dá uma idéia geral do desempenho dos dois métodos.

5.6.2 Mapa de bordas *X* Mapa de bordas + GRAS

Como nos experimentos com o uso de técnicas baseadas em cor, aqui também escolheu-se uma metodologia clássica como medida de comparação com os métodos propostos. A maneira mais tradicional de representar bordas de uma imagem é através de seu mapa

Revocação	HSV-162	HSV-162 + GRAS	Ganho (%)
0.10	0.3801	0.7796	51.2378
0.20	0.2945	0.6648	55.7010
0.30	0.2457	0.6630	62.9384
0.40	0.2265	0.6606	65.7130
0.50	0.2108	0.6589	67.9997
0.60	0.1962	0.6536	69.9816
0.70	0.1830	0.2457	25.4834
0.80	0.1718	0.2656	35.3163
0.90	0.1646	0.2085	21.0552
1.00	0.1560	0.1555	-0.3215
Média	0.2229	0.4956	55.0144

Tabela 5.4: Desempenho do histograma-162 e do histograma-162 combinado com o GRAS para 175 consultas da classe texturas da base natural.

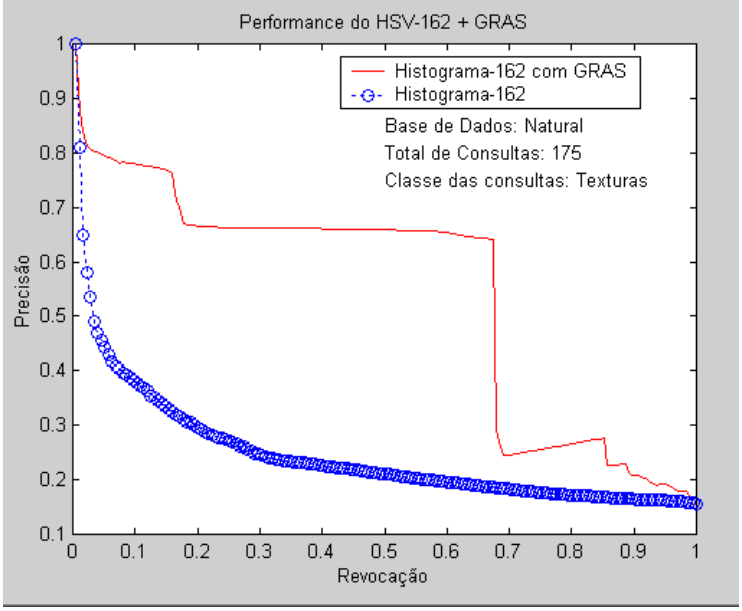


Figura 5.5: Desempenho do histograma-162 combinado com o GRAS para a classe textura da base natural. Cada uma das 175 imagens dessa classe foi tomada como consulta e confrontada com as 1200 imagens restantes da base. O gráfico é a média dessas 175 consultas.

Revocação	HSV-162	HSV-162 + GRAS	Ganho (%)
0.10	0.3375	0.5548	39.1701
0.20	0.2752	0.4985	44.7841
0.30	0.2365	0.4725	49.9386
0.40	0.2049	0.4315	52.5169
0.50	0.1808	0.3968	54.4355
0.60	0.1657	0.3539	53.1955
0.70	0.1517	0.3077	50.6986
0.80	0.1393	0.2570	45.7983
0.90	0.1256	0.2045	38.6015
1.00	0.1005	0.0975	-3.0769
Média	0.1918	0.3575	46.3547

Tabela 5.5: Desempenho do histograma-162 e do histograma-162 combinado com o GRAS para 102 consultas da classe pôr-do-sol da base natural. As 102 revocações originais foram interpoladas para as 10 mostradas na figura.

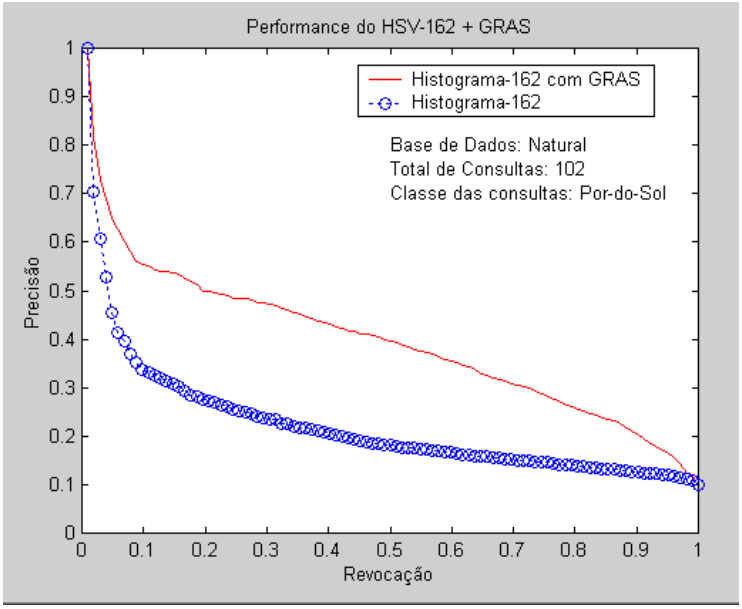


Figura 5.6: Desempenho do histograma-162 combinado com o GRAS para a classe Por-do-Sol da base natural. Cada uma das 102 imagens dessa classe foi tomada como consulta e confrontada com as 1200 imagens restantes da base. O gráfico é a média dessas 102 consultas.

de bordas. Essa técnica transforma primeiramente a imagem no espaço cromático para o espaço monocromático com 8 tons de cinza. A partir da imagem em tons de cinza, diversos filtros podem ser usados para extrair os gradientes de direção dos *pixels*. Nesta tese, foi escolhido o filtro Sobel [Gonzalez e Woods, 1992], considerando a sua simplicidade de implementação, baixo custo computacional e por ter resultados que atendem aos propósitos deste trabalho.

Uma vez calculados esses gradientes, é feita uma quantização de valores de gradientes entre zero e cento e oitenta graus para dez valores inteiros (um a 9 graus). Com esses dez valores é montado um vetor de características de dez entradas que representa a imagem de entrada.

Como na busca por cor, os experimentos foram realizados comparando-se os resultados das consultas — em termos de gráficos de Precisão x Revocação — com base em vetores de gradientes direcionais e buscas com base em vetores com gradientes direcionais combinados com 6 características extraídas pelo GRAS, as mesmas usadas para busca por cor. Assim, os vetores combinados possuem $10 + 6 = 16$ entradas.

Para a base de dados natural, observaram-se os mesmos critérios usados anteriormente. Para cada imagem foi gerado um vetor de características de 10 entradas (10 direções de bordas quantizadas) para o método clássico e 16 entradas (10 direções de bordas e 6 características do GRAS) para o método combinado. Ambos os métodos foram aplicados por classes à base natural e seu desempenho estudado. Cada classe da base natural, como já foi mencionado, possui número de membros variados, e cada um foi tomado como consulta e confrontado com as 1200 imagens da base. No final, uma média das consultas para cada classe foi calculada e as precisões e revocacões observadas.

A Tabela 5.6 mostra o desempenho do método de mapa de bordas com relação a ele mesmo combinado com o GRAS. Nota-se desta vez, uma sensível piora no desempenho do método combinado entre as revocações de 0.4 e 0.8. No entanto, o método combinado ainda tem um desempenho global melhor, média de 7%. Essa queda de desempenho do GRAS em relação ao método clássico é uma particularidade da classe que está sendo usada. Se a base possuir muitas imagens com direções bem definidas, o GRAS tende a influenciar

Revocação	Mapa de Bordas	Mapa de Bordas + GRAS	Ganho (%)
0.10	0.3301	0.4890	32.4880
0.20	0.3113	0.4172	25.3835
0.30	0.3106	0.3442	9.7341
0.40	0.3073	0.3029	-1.4526
0.50	0.3011	0.2776	-8.4654
0.60	0.2899	0.2637	-9.9355
0.70	0.2702	0.2458	-9.9491
0.80	0.2447	0.2311	-5.8849
0.90	0.2151	0.2069	-3.9642
1.00	0.1500	0.1513	0.8592
Média	0.2730	0.2930	6.7998

Tabela 5.6: Desempenho do mapa de bordas com 10 valores de direções quantizados X mapa de bordas combinado com o GRAS, para a classe carros.

pouco na busca — uma vez que captura melhor informações bem distribuídas —, caindo o seu desempenho. Pode ser exatamente o que acontece com a classe carro, devido ao excesso de imagens em ambientes abertos com gradientes de direções bem definidos. A Fig. 5.7 mostra a forma desse comportamento. Nota-se que, quando o GRAS tende a perder desempenho, o desempenho global se aproxima do método tradicional.

A seguir são mostrados as tabelas e os gráficos de desempenho de outras duas classes da base natural, elefantes (Tabela 5.7 e Fig. 5.8) e poses (Tabela 5.7 e Fig. 5.9). Ambos foram construídos usando os mesmos critérios citados anteriormente; apenas o número de consultas para cada classe varia: 96 para a classe de elefantes e 31 para a de poses.

5.6.3 Textura X Textura + GRAS

O terceiro tipo de método clássico implementado e avaliado de encontro ao GRAS, é a recuperação baseada em textura. Como foi explicado na Seção 1.1.1, existem diversas maneiras de extrair informações de textura de uma imagem; adotou-se o uso de informações da matriz de co-ocorrência [Gonzalez e Woods, 1992]. Primeiramente, quatro matrizes de co-ocorrência são calculadas para cada imagem, cada uma em uma direção diferente ($0^o, 45^o, 90^o$ e 135^o). Em seguida, os cinco descritores definidos na Seção 3.4 são calculados

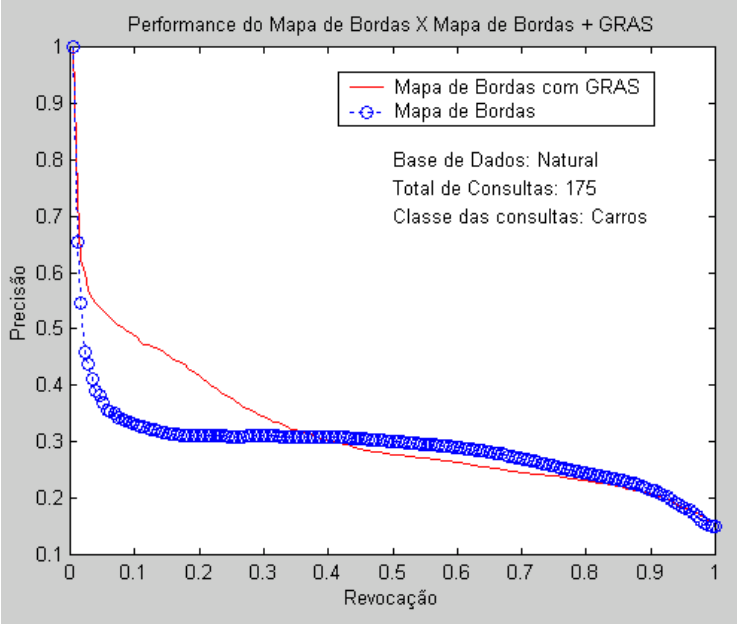


Figura 5.7: Desempenho do mapa de bordas comparado com o desempenho do mapa de bordas combinado com GRAS, para a classe carros. Quando há perda no desempenho da estratégia combinada em relação a não combinada, a perda é relativamente pequena de maneira que a estratégia combinada mantém uma média global melhor.

Revocação	Mapa de Bordas	Mapa de Bordas + GRAS	Ganho (%)
0.10	0.4402	0.4905	10.2593
0.20	0.3559	0.3956	10.0404
0.30	0.3011	0.3373	10.7257
0.40	0.2682	0.2919	8.1250
0.50	0.2357	0.2564	8.0733
0.60	0.2063	0.2188	5.7125
0.70	0.1887	0.1989	5.1181
0.80	0.1761	0.1841	4.3672
0.90	0.1584	0.1639	3.3435
1.00	0.1033	0.1086	4.8803
Média	0.2434	0.2646	8.0172

Tabela 5.7: Desempenho do mapa de bordas com 10 valores de direções quantizados X mapa de bordas combinado com o GRAS, para a classe elefante.

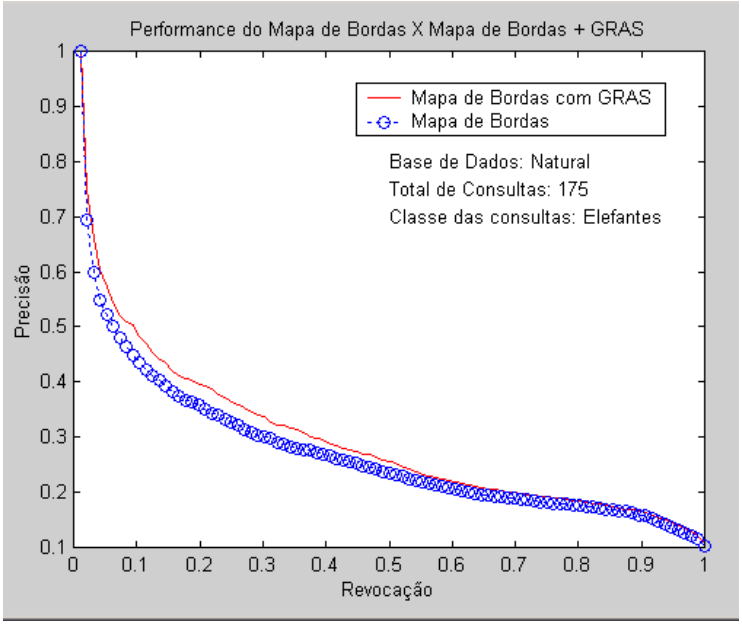


Figura 5.8: Desempenho do mapa de bordas comparado com o desempenho do mapa de bordas combinado com GRAS, para a classe elefante.

Revocação	Mapa de Bordas	Mapa de Bordas + GRAS	Ganho (%)
0.10	0.2268	0.3032	25.1896
0.20	0.1450	0.1906	23.9614
0.30	0.1398	0.1429	2.1620
0.40	0.1191	0.1237	3.7498
0.50	0.1120	0.1079	-3.7998
0.60	0.1032	0.0988	-4.4121
0.70	0.0946	0.0869	-8.8513
0.80	0.0867	0.0783	-10.7590
0.90	0.0684	0.0606	-12.8649
1.00	0.0490	0.0405	-20.9877
Média	0.1145	0.1234	7.2087

Tabela 5.8: Desempenho do mapa de bordas com 10 valores de direções quantizados X mapa de bordas combinado com o GRAS, para a classe poses.

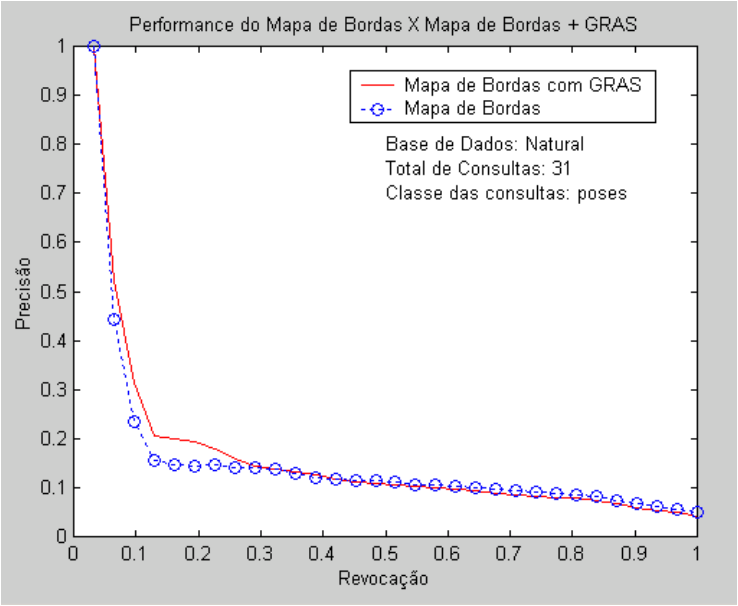


Figura 5.9: Desempenho do mapa de bordas comparado com o desempenho do mapa de bordas combinado com GRAS, para a classe poses.

para cada matriz. Assim, cada imagem da base de dados possui 4 matrizes x 5 descritores, num total de 20 pontos, que vão formar o vetor de características.

Mais uma vez, como nos experimentos anteriores, as características de textura são combinadas com as informações do GRAS e os métodos comparados com o uso da precisão e da revocação.

Assim, a Tabela 5.9 e a Fig. 5.10 são os resultados experimentais com a classe de texturas. Nesse caso, o uso do algoritmo GRAS para complementar informações melhorou o desempenho em quase 60% em média. Pode ser notado, no entanto, a grande semelhança com a Fig. 5.5. Isso pode ser indicativo de que, como as imagens de texturas usadas são ricas em pequenas regiões, os valores das características do GRAS prevalecem sobre as de textura, dando origem à gráficos similares.

Os experimentos mostrados na Tabela 5.10 exibem o desempenho do método clássico de recuperação de informações de textura em confronto com o mesmo método combinado com as características do GRAS, para a classe de 135 imagens de cenas urbanas. Para as revocações entre 0.2 e 1.0 houve uma perda de desempenho; no entanto, o ganho médio

Revocação	Textura	Textura + GRAS	Ganho (%)
0.10	0.3547	0.7214	50.8315
0.20	0.2850	0.6172	53.8183
0.30	0.2189	0.6104	64.1425
0.40	0.1708	0.6087	71.9445
0.50	0.1488	0.6140	75.7705
0.60	0.1412	0.6266	77.4654
0.70	0.1387	0.2462	43.6699
0.80	0.1381	0.2657	48.0170
0.90	0.1406	0.2092	32.8051
1.00	0.1463	0.1553	5.8106
Média	0.1883	0.4675	59.7185

Tabela 5.9: Desempenho do método baseado em textura X ele mesmo combinado com o GRAS, para a classe textura.

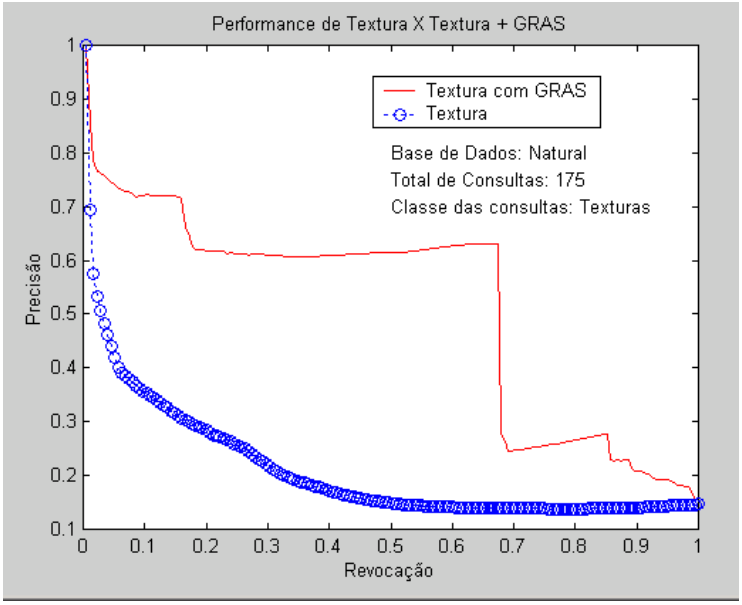


Figura 5.10: Desempenho do método baseado em texturas comparado com o desempenho dele mesmo combinado com GRAS, para a classe textura.

Revocação	Textura	Textura + GRAS	Ganho (%)
0.10	0.1506	0.1633	7.7698
0.20	0.1339	0.1276	-4.9195
0.30	0.1263	0.1290	2.1026
0.40	0.1244	0.1294	3.8428
0.50	0.1210	0.1289	6.0989
0.60	0.1200	0.1302	7.8030
0.70	0.1194	0.1297	7.9476
0.80	0.1186	0.1272	6.7827
0.90	0.1201	0.1239	3.0685
1.00	0.1161	0.1157	-0.3610
Média	0.1250	0.1305	4.1717

Tabela 5.10: Desempenho do método baseado em textura X ele mesmo combinado com o GRAS, para a classe urbanas.

global ficou em torno de 4.17%. Esse tipo de imagem, dado a sua diversidade, está entre os mais difíceis de se obter ganho. No entanto, a Fig. 5.11 mostra que o comportamento geral é competitivo com o método clássico.

A Tabela 5.11 mostra o desempenho do método de busca por textura comparado com esse mesmo método combinado com características do GRAS, para a classe de por-do-sol (com 102 imagens) da base natural. Para essa classe, o método combinado mostrou-se mais eficiente a partir de 20% de revocação, a partir daí a combinação dos dois métodos ganha sempre. A Fig. 5.12 mostra graficamente esse comportamento.

5.7 Métodos propostos

5.7.1 HSV + ASL

Nesta seção, são exibidos e analisados os experimentos com o uso da Análise Semântica Latente 4.2.

Como nos experimentos anteriores, esse método foi testado por classes. A primeira tomada como exemplo foi a de poses, a qual possui cerca de 31 imagens de faces humanas. Primeiramente, observa-se o desempenho do método de histograma-162 contra o desempe-

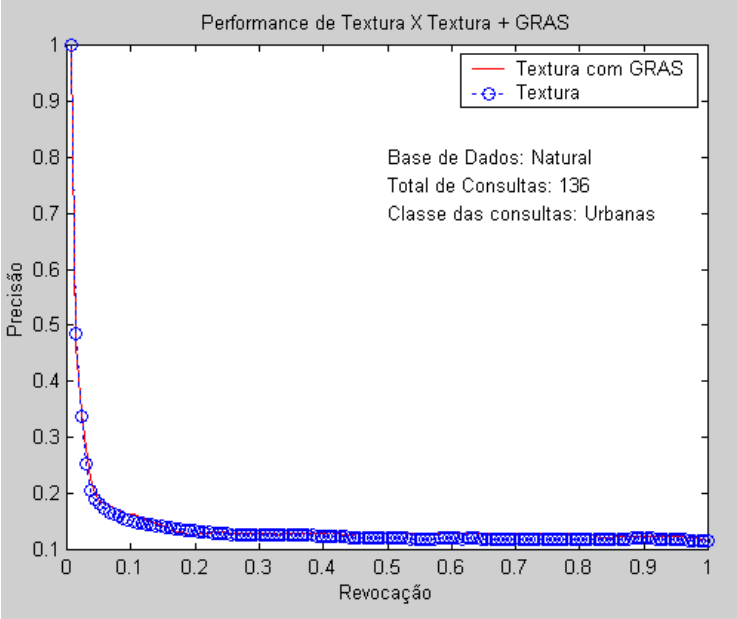


Figura 5.11: Desempenho do método baseado em texturas comparado com o desempenho dele mesmo combinado com GRAS, para a classe de cenas urbanas.

Revocação	Textura	Textura + GRAS	Ganho (%)
0.10	0.3288	0.2612	-25.8586
0.20	0.1854	0.2391	22.4716
0.30	0.1495	0.2308	35.2239
0.40	0.1278	0.2270	43.7244
0.50	0.1164	0.2235	47.8987
0.60	0.1084	0.2211	50.9564
0.70	0.1024	0.2114	51.5741
0.80	0.1004	0.1983	49.3482
0.90	0.0960	0.1708	43.7649
1.00	0.0905	0.0944	4.1564
Média	0.1406	0.2078	32.3449

Tabela 5.11: Desempenho do método baseado em textura X ele mesmo combinado com o GRAS, para a classe de Por-do-Sol.

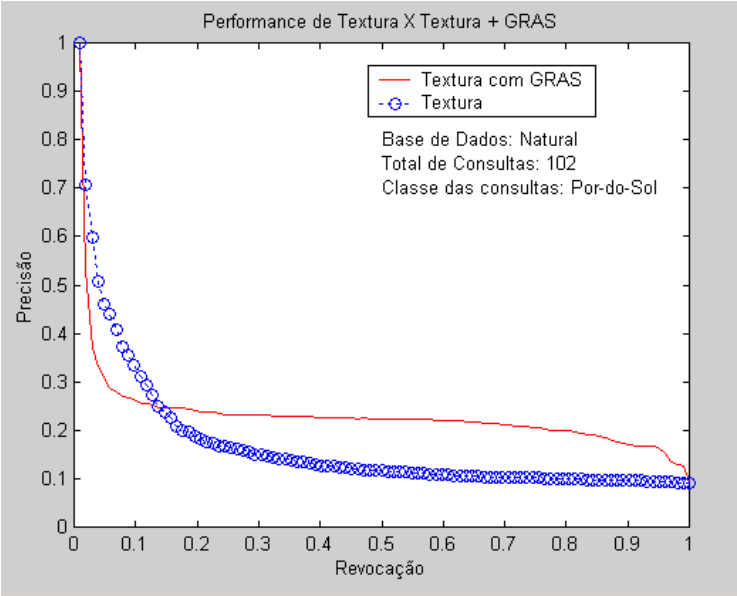


Figura 5.12: Desempenho do método baseado em texturas comparado com o desempenho dele mesmo combinado com GRAS, para a classe pôr-do-sol.

nho do mesmo método após a transformação do espaço latente. Pela Tabela 5.12, o ganho com a mudança de espaço foi sutil nessa classe; isso pode ser confirmado observando-se o comportamento global do gráfico, mostrado na Fig. 5.13. Essa mudança, apesar de ser pequena comparada com ganho obtido pelo GRAS, indica um aumento da informação semântica na consulta, fornecida pelo espaço latente.

A Tabela 5.13 exibe os resultados dos experimentos com a classe de setas. Essa classe consta de 44 imagens artificiais de setas, sendo portanto, imagens bastante específicas. Como essa classe é a única artificial de toda a base, ela é relativamente fácil de recuperar, o que explica o bom desempenho, tanto do método que usa somente histograma-162 quanto após a transformação latente. Nesse último caso, como ocorreu na classe de poses, o ganho foi em torno de 5%. A Fig. 5.14 mostra o comportamento das curvas. Nota-se no entanto, que, ao contrário da classe de setas, nesse caso há ganho em todos os pontos de revocação.

O desempenho do método que usa ASL combinado com histograma-162 é mostrado na Tabela 5.14 e na Fig. 5.15. A classe usada aqui foi a de carros, a qual possui 175 imagens, todas usadas como consulta e, como em todos os experimentos dessa tese, representada

Revocação	HAV	HSV + ASL	Ganho (%)
0.10	0.3054	0.2927	-4.3224
0.20	0.1680	0.1881	10.6572
0.30	0.1393	0.1582	11.9591
0.40	0.1179	0.1326	11.1011
0.50	0.1024	0.1092	6.2733
0.60	0.0917	0.0963	4.8408
0.70	0.0765	0.0790	3.1913
0.80	0.0639	0.0660	3.2288
0.90	0.0507	0.0501	-1.2808
1.00	0.0363	0.0354	-2.5321
Média	0.1152	0.1208	4.6092

Tabela 5.12: Desempenho do histograma-162 *X* histograma-162 combinado com o ASL. Média de 31 consultas com imagens de faces humanas (classe poses).

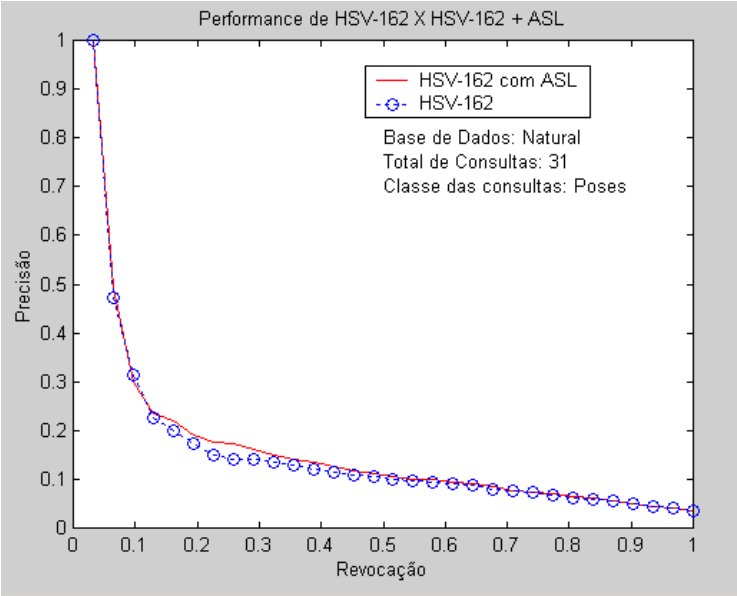


Figura 5.13: Desempenho do histograma-162 comparado com o desempenho do histograma-162 combinado com ASL. Os valores para a construção desse gráfico foram tomados a partir da média das 31 imagens da classe.

Revocação	HSV	HSV + ASL	Ganho (%)
0.10	0.9100	0.9101	0.0174
0.20	0.8787	0.8900	1.2699
0.30	0.8588	0.8868	3.1542
0.40	0.7948	0.8748	9.1494
0.50	0.7775	0.8462	8.1168
0.60	0.7683	0.8227	6.6186
0.70	0.7731	0.8028	3.6960
0.80	0.7592	0.7921	4.1587
0.90	0.5504	0.5744	4.1773
1.00	0.0499	0.0502	0.6429
Média	0.7121	0.7450	4.4233

Tabela 5.13: Desempenho do histograma-162 X histograma-162 combinado com o ASL, para a classe de Setas.

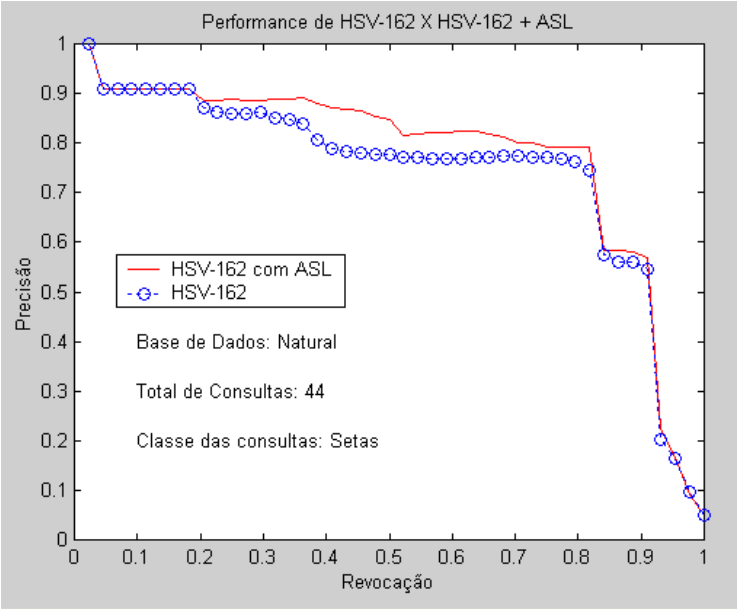


Figura 5.14: Desempenho do histograma-162 comparado com o desempenho do histograma-162 combinado com ASL, para a classe Setas.

Revocação	HSV	HSV + ASL	Ganho (%)
0.10	0.3421	0.3612	5.2942
0.20	0.3037	0.3157	3.8004
0.30	0.2739	0.2814	2.6654
0.40	0.2476	0.2528	2.0783
0.50	0.2285	0.2318	1.4212
0.60	0.2114	0.2135	0.9965
0.70	0.1983	0.1988	0.2756
0.80	0.1876	0.1872	-0.2372
0.90	0.1779	0.1784	0.2943
1.00	0.1578	0.1581	0.2082
Média	0.2329	0.2379	2.1126

Tabela 5.14: Desempenho do histograma-162 X histograma-162 combinado com o ASL, para a classe de Carros.

pela média. Esse caso é notável pela pouquíssima diferença que faz a transformação latente no espaço vetorial. A classe carro, como já foi mencionado, é uma das que possui maior dificuldade na classificação manual. Isso se reflete no espaço latente, baixando o seu desempenho.

5.7.2 HSV + Mapa de Bordas + Textura: Modelo Bayesiano

Nesta seção serão apresentados os experimentos com o uso do modelo bayesiano proposto na Seção 4.3. Esse modelo pode ser usado junto ou em separado com as metodologias apresentadas acima. No entanto, só serão mostrados os testes combinando os três modelos clássicos.

Primeiramente, como já é um resultado conhecido que agrupamento da base de dados melhora o desempenho dos sistemas, todas as bases foram agrupadas previamente usando as técnicas apresentadas na Seção **??**.

O primeiro experimento que será analisado é sobre a classe da base natural felinos. Seus resultados podem ser observados na Tabela 5.15, e o comportamento dos gráficos na Fig. 5.16. A sexta coluna da tabela representa o ganho da rede em relação ao método (cor, borda ou textura) que obteve o melhor desempenho entre os três. Assim, ora o ganho (ou

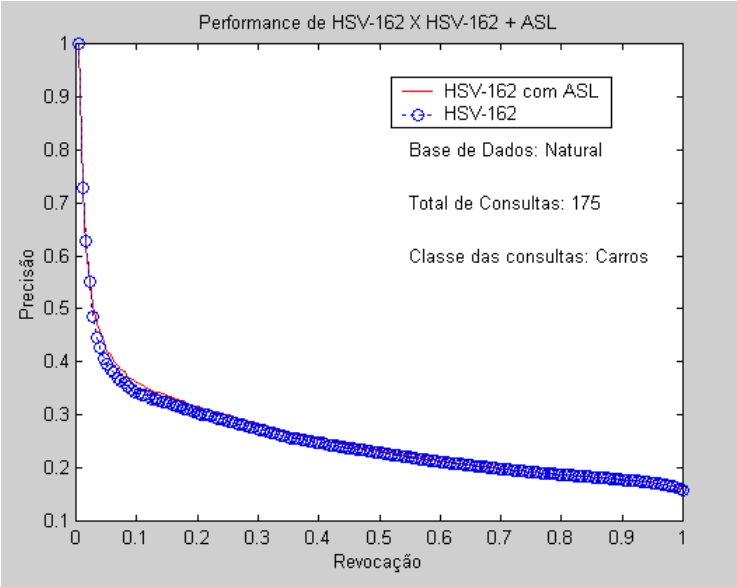


Figura 5.15: Desempenho do histograma-162 comparado com o desempenho do histograma-162 combinado com ASL, para a classe Carro.

perda) da rede pode ser tomado com relação ao histograma-162, ora com relação ao mapa de bordas e ora com relação ao método que usa textura, dependendo da classe usada.

A primeira observação que pode trivialmente ser feita é com respeito a ordem de desempenho dos quatro métodos. O pior desempenho, nessa classe, foi para o método que usa textura para recuperação, seguido do histograma-162, mapa de bordas e o modelo bayesiano. Na verdade, pelo gráfico pode ser observado que, para revocações acima de 55% (acima de 60 imagens relevantes recuperadas) o desempenho do modelo bayesiano passa para o segundo lugar, no entanto, ganha na média. Para revocações mais baixas, quando o usuário requer um número pequeno de imagens recuperadas, o ganho da rede é em torno de 6% a 17% de ganho. Nessa classe, o ganho da rede bayesiana foi tomado em relação ao método de mapa de bordas, por este ter tido o melhor desempenho para esta classe. O gráfico é a média de 105 buscas de imagens de felinos.

A segunda classe mostrada é a classe de aviões. O desempenho do modelo bayesiano é mostrado na Tabela 5.16 e na Fig. 5.17. Nesse caso, o ganho da rede foi medido em relação ao método histograma-162, que obteve melhor desempenho entre os três métodos clássicos. Tanto o gráfico quanto a tabela é a média de 101 buscas com imagens de aviões.

Revocação	HSV	Mapa de Bordas	Textura	Bayesiano	Ganho (%)
0.10	0.4815	0.5198	0.1730	0.6270	17.0827
0.20	0.3854	0.4557	0.1265	0.5173	11.9080
0.30	0.3349	0.4181	0.1167	0.4582	8.7407
0.40	0.3000	0.3922	0.1117	0.4170	5.9472
0.50	0.2744	0.3758	0.1109	0.3791	0.8705
0.60	0.2474	0.3584	0.1125	0.3457	-3.6737
0.70	0.2215	0.3367	0.1101	0.3036	10.8860
0.80	0.1880	0.3132	0.1072	0.2594	-20.7402
0.90	0.1497	0.2808	0.1041	0.2129	-31.8694
1.00	0.1083	0.1221	0.0963	0.1467	16.7689
Média	0.2691	0.3573	0.1169	0.3667	2.5649

Tabela 5.15: Desempenho do modelo bayesiano proposto com relação aos três métodos clássicos estudados. A sexta coluna mostra o ganho (ou perda) em relação ao método demapa de bordas. A classe usada aqui é a classe felinos.

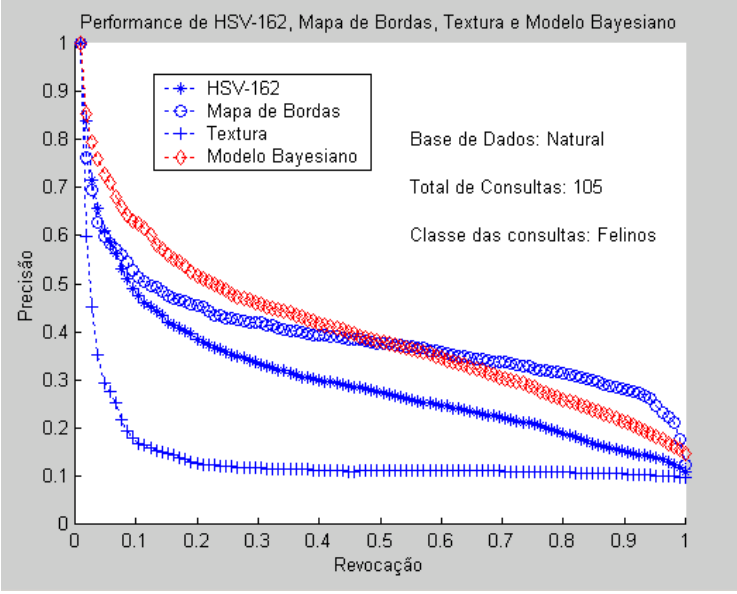


Figura 5.16: Desempenho do modelo bayesiano em relação aos métodos clássicos para a classe felinos.

Revocação	HSV	Mapa de Bordas	Textura	Bayesiano	Ganho (%)
0.10	0.6083	0.5023	0.3695	0.7036	13.5499
0.20	0.5567	0.4324	0.3311	0.6162	9.6530
0.30	0.5274	0.3709	0.3054	0.5618	6.1210
0.40	0.4949	0.3204	0.2705	0.5223	5.2533
0.50	0.4098	0.2716	0.2404	0.4546	9.8658
0.60	0.2811	0.2100	0.2155	0.3538	20.563
0.70	0.2231	0.1532	0.1906	0.2662	16.191
0.80	0.1267	0.1242	0.1559	0.1902	33.357
0.90	0.1010	0.1034	0.0924	0.1351	25.242
1.00	0.0880	0.0913	0.0895	0.0944	6.7797
Média	0.3417	0.2580	0.2261	0.3898	33.8283

Tabela 5.16: Desempenho do modelo bayesiano na classe aviões.

Revocação	HSV	Mapa de Bordas	Textura	Bayesiano	Ganho (%)
0.10	0.3831	0.2851	0.2487	0.4743	39.8912
0.20	0.2838	0.2640	0.1602	0.3727	29.1624
0.30	0.2390	0.2432	0.1371	0.3210	24.2477
0.40	0.2092	0.2292	0.1169	0.2853	19.6762
0.50	0.1872	0.2117	0.1089	0.2600	18.5769
0.60	0.1704	0.2001	0.1053	0.2344	14.6404
0.70	0.1529	0.1816	0.1043	0.2076	12.4880
0.80	0.1406	0.1623	0.1026	0.1790	9.3509
0.90	0.1230	0.1325	0.0993	0.1424	6.9542
1.00	0.0987	0.0883	0.0908	0.0905	2.4309
Média	0.1988	0.1998	0.1274	0.2567	22.1744

Tabela 5.17: Desempenho do modelo bayesiano na classe pôr-do-sol.

A Tabela 5.17 e a Fig. 5.18 mostram o desempenho da rede bayesiana na classe pôr-do-sol. Também neste caso, como no anterior (classe de aviões), a rede bayesiana foi melhor em todos os pontos de revocação. Seus ganhos foram obtidos com relação ao método baseado em mapa de bordas, e são a média de 102 buscas com imagens de pôr-do-sol.

Finalmente, é interessante observar os resultados do uso da rede bayesiana na classe de fotografias de pessoas. São apenas 31 imagens de pessoas em fotos de rostos. O ganho é em relação ao método de histograma-162. No entanto, devido a pouca representatividade

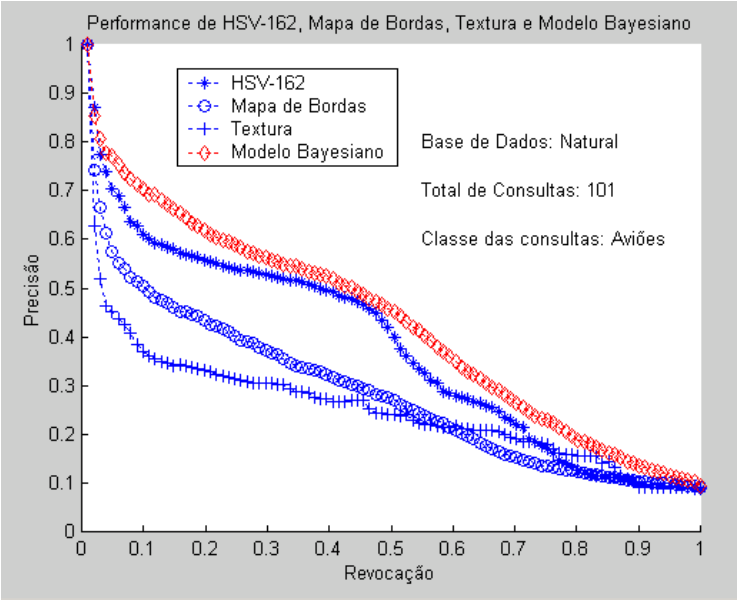


Figura 5.17: Desempenho do modelo bayesiano na classe aviões. Pode-se notar que, nesta classe, a rede é melhor para todos os pontos de revocação.

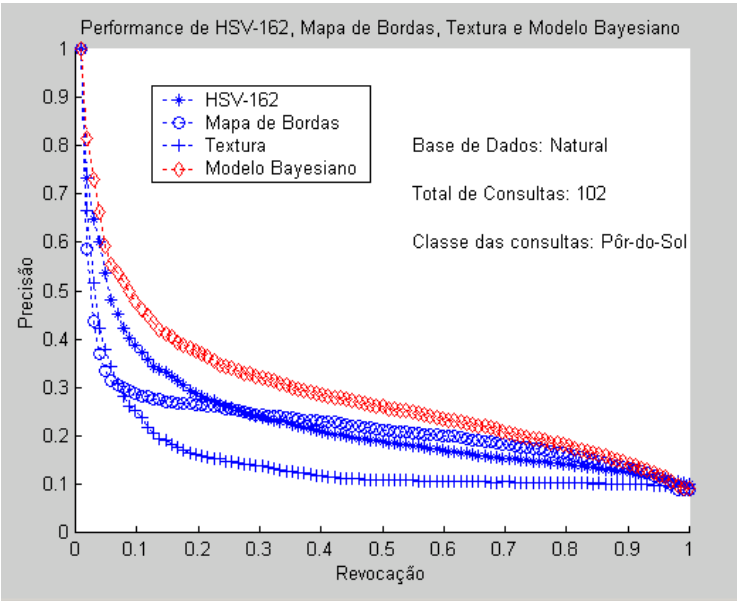


Figura 5.18: Desempenho do modelo bayesiano na classe pôr-do-sol.

Revocação	HSV	Mapa de Bordas	Textura	Bayesiano	Ganho (%)
0.10	0.3769	0.3350	0.1944	0.5881	35.9067
0.20	0.2007	0.2301	0.1184	0.3440	41.6638
0.30	0.1606	0.1297	0.0948	0.2388	32.7681
0.40	0.1174	0.0681	0.0669	0.1147	-2.3183
0.50	0.0905	0.0586	0.0441	0.0771	-17.368
0.60	0.0698	0.0458	0.0349	0.0560	-24.562
0.70	0.0554	0.0336	0.0332	0.0445	-24.746
0.80	0.0365	0.0293	0.0348	0.0352	-3.8090
0.90	0.0253	0.0246	0.0322	0.0251	-0.9171
1.00	0.0260	0.0260	0.0304	0.0260	0
Média	0.1159	0.0981	0.0684	0.1550	36.7024

Tabela 5.18: Desempenho do modelo bayesiano na classe de fotografias de pessoas.

da base, poucas conclusões seguras podem ser tiradas. Uma delas é que o modelo teve bom desempenho para pequenas revocaçãoes.

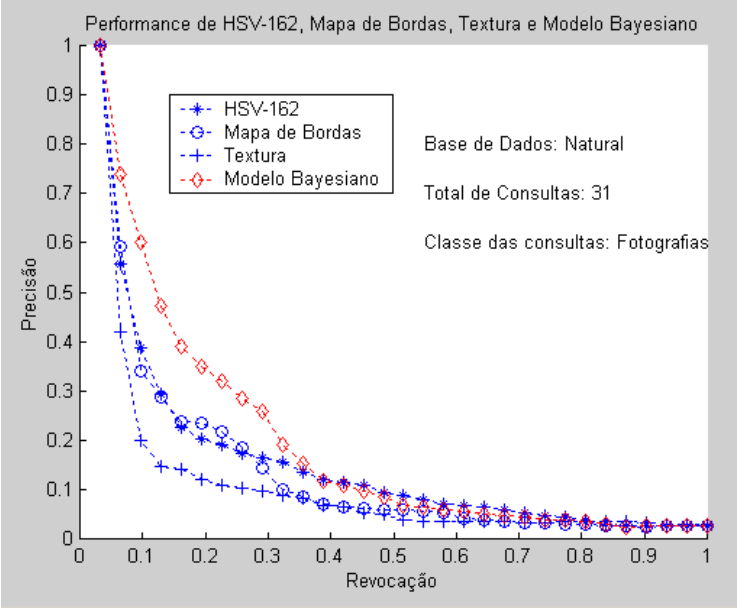


Figura 5.19: Desempenho do modelobayesiano na classe de fotografia de pessoas.

Capítulo 6

Conclusões

Neste trabalho, foram apresentadas três novas estratégias como soluções para as etapas de extração de características e ordenação de conjuntos-resposta em RIBC. Cada uma atua de forma diferente. A primeira delas é o algoritmo denominado de GRAS, que se baseia em informações espaciais da imagem como estratégia para adicionar mais informações às utilizadas nos métodos tradicionais, tais como cor, borda e textura. A segunda estratégia, denominada ASL, usa a decomposição em valores singulares como forma de transformar o espaço vetorial em um espaço de informações que co-ocorrem, o espaço latente. Ao contrário do GRAS, que adiciona informações a cada vetor de características da base, a ASL é usada apenas como uma etapa anterior à busca, transformando o espaço de busca em um outro espaço, o espaço latente, onde as informações são melhor distribuídas. A terceira estratégia, baseada em redes bayesianas, adota uma abordagem bem distinta das anteriores que, ao invés de adicionar mais informações ao espaço vetorial, ou transformá-lo para um outro espaço, atua diretamente na ordenação final dos vetores, combinando busca por cor, forma e textura.

O aumento de informações da base de dados, com o intuito de melhorar o processo de reconhecimento é, no entanto, ambíguo. Isso porque, ao se aumentar a quantidade de informação mínima que um sistema, biológico ou artificial, precisa para o processo de reconhecer um objeto ou cena, pode-se estar gerando informações redundantes, que

mais contribuem para gerar ambigüidade nesse processo do que realizar a tarefa com mais precisão. Em outras palavras, num processo de reconhecimento, mais informação não significa maior certeza. Por exemplo, num sistema construído para reconhecer padrões de plantas como orquídeas, a grande variedade de cores diferentes que podem existir sugere a implementação de detectores de formas, ignorando a cor ou a textura. Se a informação de cor for adicionada, isso pode gerar um colapso no processo de detecção. Da mesma forma, um sistema para detecção de cenas naturais que contenham céu, mar, vegetação, etc., intuitivamente deve basear-se em informações de cores ou textura, uma vez que a grande variedade de formas que podem existir mais contribuem para dificultar o reconhecimento do que realizá-lo.

No entanto, medir a quantidade de informação necessária que um sistema de reconhecimento precisa utilizar para realizar suas tarefas com precisão é um processo difícil, para não dizer impossível de realizar. Adicionar informações pode tanto influenciar negativamente quanto positivamente na tarefa final que é reconhecer um objeto ou cena.

Assim, quando se usa informações espaciais para adicionar mais informações aos vetores de características de uma imagem, pode-se estar dificultando a busca e não necessariamente melhorando. No entanto, o resultado, positivo ou negativo, vai depender de diversos fatores como: o tipo de imagem que se está utilizando, a complexidade da cena, a quantidade de informação que se está adicionando e o tipo de informação. Uma cena natural, como as da base natural utilizada neste trabalho, por exemplo, geralmente podem ser reconhecidas através de elementos tanto de cor, forma e textura quanto do relacionamento espacial entre essas as regiões que compõem a imagem. Uma cena de elefantes possui geralmente forma (a dos elefantes), cor (do céu ou da vegetação), e textura (da vegetação ou da pele dos elefantes) e também possui elementos que se localizam em uma determinada disposição (o céu fica na parte de cima, depois a vegetação, etc.). Esses elementos geram informações para o GRAS e podem ser sutilmente adicionados aos vetores de características como aumento de informações. A qualidade do reconhecimento, entretanto, só poderá ser medida estatisticamente e, vai depender, como foi explicado, de todos esses fatores. Esse é um processo observado mais claramente nas base natural do que na base artificial, uma

vez que esta última possui geralmente menos regiões do que a primeira.

No caso das redes bayesianas, por ser um sistema que combina separadamente atributos de cor, forma e textura, trata de maneira diferente essas informações. Não se está aqui combinando no mesmo vetor de características informações diferentes, como no caso do GRAS, mas combinando os resultados individuais. No entanto, para a estratégia adotada pelo GRAS, para as bases de dados, os resultados foram claramente melhores. Quando os atributos envolvidos (cor, forma e textura) geram resultados próximos, as informações são combinadas de maneira que o resultado final seja melhor do que o melhor dos atributos. Por outro lado, quando pelo menos um atributo gera resultados muito abaixo do esperado, ele tende a trazer o resultado final para baixo, ficando este, geralmente, pior do que o melhor atributo, gerando um resultado inferior ao caso sem o uso da rede. Todos esses dados são confirmados pelos experimentos.

A primeira bateria de testes comparou a aplicação das técnicas clássicas de histograma-162, mapa de bordas e textura, combinadas com as características extraídas pelo GRAS. Esses testes foram conduzidos em duas bases de dados. Na base Columbia, uma base controlada de imagens de objetos com fundo homogêneo preto, os resultados combinados com o GRAS não foram satisfatórios. Por sua vez, quando os experimentos foram conduzidos sobre a base natural, obtiveram-se resultados melhores, o que é indício da influência positiva do GRAS para esses casos. A explicação para isso vem das características de cada base e do objetivo do GRAS. Para a base controlada, são geradas poucas regiões de uma forma mais definida; para a base natural, é normal que seja gerado um número bem maior de regiões. Como o GRAS trabalha sobre regiões, é normal que se obtenha melhores resultados em bases desse tipo, como a base natural. Logo, pode-se concluir que esse algoritmo é mais útil para sistemas de recuperação de imagens naturais do que para sistemas puramente de reconhecimento de objetos, como sistemas de inteligência artificial.

Em seguida, foram realizados os testes com a Análise Semântica Latente. A condução dos experimentos foi semelhante. Compararam-se as técnicas de histogramas-162, mapa de bordas e textura, com e sem a transformação do espaço latente. No entanto, os resultados da base Columbia foram também inferiores aos da base natural. A explicação para isso

repousa nos mesmos argumentos discutidos para o GRAS. A ALS é uma técnica que se prevale de atributos globais da cena para gerar informações que serão acrescentadas ao espaço de busca. No entanto, as imagens da Columbia são artificiais, com fundo homogêneo, portanto, produzem poucas regiões devido seus contextos simples. Nesse caso, a ASL tem pouca informação a acrescentar.

Para a base de dados natural, mais uma vez obteve-se ganho. Esse ganho foi também pequeno, considerando a performance do GRAS. Se for retirada uma média, pode-se chegar a um valor em torno de 5%. Não é muito, mas duas observações primordiais podem ser feitas. A primeira é que a ASL usa informações implícitas, e que não podem ser capturadas apenas com esses métodos tradicionais, indicando um "caminho a ser seguido", caso deseje-se implementar esse tipo de informação contextual. Investigar novas maneiras do uso da ASL, com o intuito de extrair informações semânticas de uma base vetorial, é objeto de estudos modernos [Kintsch, 2001]. A segunda observação é que, pelo mesmo motivo do GRAS, devido ao maior número de informações nas imagens da base natural, é de se esperar que a ASL tenha um desempenho melhor, uma vez que trabalha em cima de informações contextuais.

Nos experimentos usando o modelo bayesiano proposto, os resultados mostram que, quando se usa cor, forma e textura ao mesmo tempo, para recuperação de informação visual, o modelo, em média, produz um desempenho que é melhor do que o desempenho da melhor das três técnicas, caso fosse usada isoladamente. Os resultados mais promissores foram obtidos quando o desempenho dos métodos baseados em cor, borda e textura são próximos. Por outro lado, quando algum método tem um desempenho muito ruim, a rede tende a sofrer uma queda de performance e geralmente perde em desempenho para o melhor dos três métodos. Pode-se concluir então que a rede é mais útil quando a classe envolvida pode ser recuperada por cor, borda, ou textura, indiferentemente. Por exemplo, a classe de imagens de textura da base natural, é muito específica, fazer recuperação nela com base em cor ou borda, gera resultados piores com relação a métodos baseados unicamente em textura. Por esse motivo, a rede bayesiana proposta não obteve resultado satisfatório para esta classe.

Problema	perda de informações primitivas	perda de informações contextuais	perdas ao combinar características
Solução	aumento de informações espaciais	modelar interações das primitivas	combinar melhor as características
Proposta	GRAS	ASL	modelo bayesiano
Ganho	5 a 60%	3 a 5%	3 a 36%

Tabela 6.1: Resumo final dos métodos apresentados e estudados.

Finalmente, todos os novos métodos apresentados aqui — GRAS, ASL e rede bayesiana — melhoram em maior ou menor grau o desempenho de sistemas RIBC. A conclusão final do trabalho pode ser resumida na Tabela 6.1

Bibliografia

[Baesa-Yates e Ribeiro-Neto, 1999] Baesa-Yates, R. and Ribeiro-Neto, B. (1999). *Moderm Information Retrieval*. Addison Wesley, Harlow.

[Balakrishnan, 1997] Balakrishnan, V. K. (1997). *Theory and Problems of Graph Theory*. McGraw-Hill.

[Beckmann et al., 1990] Beckmann, N., Kriegel, H. P., Schneider, R., and Seeger, S. (1990). The R^*-tree: an efficient and robust access method for points and rectangles. In *Proc. ACM SIGMOD*, pages 322–331.

[Belongie et al., 1998] Belongie, S., Carson, C., Greenspan, H., and Malik, J. (1998). Color and texture-based image segmentation using EM and its application to content-based image retrieval. In *Proceedings of IEEE International Conference on Computer Vision (ICCV-98)*, pages 675–682, Bombay, India.

[Beltrami, 1873] Beltrami, E. (1873). Sulle funzioni bilincari. *Giornale di Matematiche ed uso Degli Studenti Delle Universitá*, 11:98–106.

[Biederman, 1987] Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147.

[Bregler e Malik, 1997] Bregler, C. and Malik, J. (1997). Learning appearance based models: mixtures of second moment experts. In M. C. Mozer et al. (Eds.), editor, *Advances in Neural Information Processing Systems 9*, pages 845–851. MIT Press.

[Brooks, 1983] Brooks, R. A. (1983). Model-based three-dimensional interpretation of two-dimensional images. *IEEE Transactions on Patterns Analysis and Machine Intelligence*, 5(2):140–150.

[Buijs e Lew, 1999] Buijs, J. M. and Lew, M. S. (1999). Visual learning of simple semantics in Image Space, VISUAL'99: Third International Conference on Visual Information and Information Systems. *Lecture Notes in Computer Science*, 1614:131–138.

[Campbell et al., 1997] Campbell, N. W., William, P. J. M., and Barry, T. T. (1997). Interpreting image database by region classification. *Pattern Recognition*, 30(4):555–567.

[Chun, 2000] Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, 4(5):170–177.

[Columbia, 2001] Columbia (2001). Department of Computer Science. Columbia University, http://www.cs.columbia.edu/CAVE/research/softlib/coil-100.html.

[Cormem et al., 1997] Cormem, T. H., Leiserson, C. E., and Rivest, R. L. (1997). *Introduction to Algorithms*. McGraw-Hill.

[Deewester et al., 1990] Deewester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Hashman, R. (1990). Indexing by Latent semantic Analysis. *Journal of the American Society for Information Science*, 41.

[Draper et al., 1989] Draper, B. A., Collins, R., Brolio, J., Hanson, A., and Riseman, E. (1989). The SCHEMA system. *International Journal on Computer Vision*, 2:209–250.

[Drimbaream e Whelan, 2001] Drimbaream, A. and Whelan, P. F. (2001). Experiments in colour texture analysis. *Patterns Recognition Letters*, 22:1161–1167.

[Durand et al., 1999] Durand, G., Thienot, C., and Faudemay, P. (1999). Extraction of composite visual objects from audiovisual materials. In S. Panchanathan et. al (Eds.), editor, *Multimedia Storage and Archiving Systems IV, Proc. SPIE 3846*, pages 194–203.

[Eakins, 2002] Eakins, J. P. (2002). Towards intelligent image retrieval. *Pattern Recognition*, 35:3–14.

[Eakins e Graham, 1999] Eakins, J. P. and Graham, M. E. (1999). Content-based image retrieval. Available online at http://www.unn.ac.uk/iidr/CBIR/report.html. JISC Technology Applications Programme Report 39.

[Flickner, 1995] Flickner, M. (1995). Query by images and video content: The QBIC system. *Special Issue on Content-Based Image Retrieval Systems*, 28(9):23–32.

[Forsyth et al., 1997] Forsyth, D. A., Malik, J., Fleck, M. M., Greespan, H., Leung, T., Belongie, S., Carson, C., and Bregler, C. (1997). Finding pictures of objects in large collections of images. In *Digital Image Access and Retrieval: 1996 Clinic on Library Applications of Data Processing Graduate School of Library and Information Scienc*, pages 118–139, University of Illinois at Urbana-Champaig.

[Fournier et al., 2001] Fournier, J., Cord, M., and Philipp-Foliguet, S. (2001). RETIN: A Content-Based Image Indexing and Retrieval System. *Pattern Analysis & Applications*, 4:153—173.

[Fuertes et al., 2001] Fuertes, J. M., Lucena, M., de la Banca, N. P., and Chamorro-Martínez, J. (2001). A scheme of colour image retrieval from databases. *Pttern Recognition Letters*, 22:323–337.

[Golub e Loan, 1989] Golub, G. H. and Loan, C. F. (1989). *Matrix Computations (2nd ed.)*. Johns Hopkins Univeristy Press, Baltimore, Maryland.

[Gonzalez e Woods, 1992] Gonzalez, R. C. and Woods, R. E. (1992). *Digital Image Processing*. Addison Wesley.

[Greenberg, 2001] Greenberg, J. (2001). Optimal query expansion (QE) processing methods with semantically encoded structural thesauri terminology. *Journal of the American Society for Information Science*, 52(6):487–498.

[Gudivada e Raghavan, 1995] Gudivada, V. N. and Raghavan, V. V. (1995). Special Issue on Content-Based Image Retrieval Systems. *Computer.*, 28(9):119–141.

[Hermes et al., 1995] Hermes, T., Klauck, C., KreiB, J., and Zhang, J. (1995). Image retrieval for information syetems. In W.R. Niblack and R. C. Jain (Eds., editor, *Storage and Retrieval for Image and Video Databases III*, pages 394–405. Proc. SPIE 240.

[Hirata e Kato, 1992] Hirata, K. and Kato, T. (1992). Query by visual example-content based image retrieval. In *In: Advances in Database Technology (EDBT'92)*, pages 56–71.

[Jacob et al., 1995] Jacob, C. E., Finkelstein, A., and Salesin, D. H. (1995). Fast multiresolution image query. In *In: Proceedings of the 1995 ACM SIGGRAPH*.

[Johnson e Wichern, 1998] Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis (4nd ed.)*. Prentice Hall.

[Jordan, 1874] Jordan, C. (1874). Mémoire sur les formes bilinéaires. *Journal de Mathématiques Pures et Appliquées. Deuxiéme Série*, 19:35–54.

[Kintsch, 2001] Kintsch, W. (2001). Prediction. *Cognitive Science*, 25:173–202.

[Lee e Kim, 2001] Lee, D. and Kim, H. (2001). A fast content-based indexing and retrieval technique by the shape information in large image database. *The Journal of Systems and Software*, 56:165–182.

[Leung e Malik, 1999] Leung, T. and Malik, J. (1999). Recognizing surfaces using three-dimensional textons. In *Seventh IEEE International Conference on Computer Vision (ICCV-99)*, volume 2, pages 1010–1017, Corfu, Greece.

[Lipson et al., 1997] Lipson, P., Grimson, E., and Sinha, P. (1997). Configuration-based scene classification and image indexing. In *Proceedings of IEEE Conference on Computer Visiob and Pattern Recognition (CVPR-97)*, pages 1007–1013, Puerto Rico.

[Manjunath e Ma, 1996] Manjunath, B. S. and Ma, W. Y. (1996). Texture features for browsing and retrieval of large image data. *IEEE Transactions on Pattern Analises and Machine Intelligence*, 18:837–842.

[Marr, 1982] Marr, D. (1982). *Vision*. San Francisco: Freeman.

[Marsicoli et al., 1997] Marsicoli, M., Cinque, L., and Levialdi, S. (1997). Indexing pictorial documents by their content: a survey of current techniques. *Image and Vision Computing.*, 15:119–141.

[Martinez e Serra, 1999] Martinez, A. and Serra, J. R. (1999). Semantic Access to a Database of Images: an approach to object-related image retrieval. In *Proceedings of IEEE Multimedia Systems (ICMCS)*, pages 624–629, Florence, Italy.

[Matsuyama e Hwang, 1990] Matsuyama, T. and Hwang, V. (1990). *SIGMA: a knowledge-based aerial image understanding system*. Plenum, New York.

[Nastar et al., 1998] Nastar, C., Mitschke, M., and Meilhac, C. (1998). Efficient query refinement for image retrieval. In *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'98)*, pages 547–552.

[Neiser, 1967] Neiser, U. (1967). *Cognitive Psychology*. New York: Appleton-Century-Crofts.

[Niblack, 1993] Niblack, W. (1993). The QBIC project: querying images by content using color, texture and shape. In *Proc. Storage and Retrieval for Image and Video Data Bases, SPIE*, volume 1, pages 173–187, Bellingham, WA.

[Oliva et al., 1999] Oliva, A., Torralba, A. B., Guerin-Dugue, A., and Herault, J. (1999). Global semantic classification of scenes using power spectrum templates. In *The Challenge of Image Retrieval*. CIR-99, Newcastle upon Tyne, UK.

[Ooi et al., 1998] Ooi, B. C., Tan, K. L., Chua, T. S., and Hsu, W. (1998). Fast image retrieval using color-spatial information. *The VLDB Journal*, 7(2):115–128.

[Paek et al., 1999] Paek, S., Sable, C. L., Hatzivassiloglou, V., Jaimes, A., Schiffman, B. H., Chang, S.-F., and McKeown, K. R. (1999). Integration of visual and text-based approaches for the content labeling and classification of photographs. In *ACM SIGIR'99 Workshop on Multimedia Indexing and Retrieval*, Berkeley, CA.

[Palmer, 1975] Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition*, 3(5):519–526.

[Pearl, 1988] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmman, Los Autos, California.

[Puzica et al., 1997] Puzica, J., Hoffmann, T., and Buhman, J. M. (1997). Non-Parametric Similarity Measuress for Unsupervised Texture Segmentation and Image Retrieval. In *IEEE Conference on Computer Vision and Patterns Recognition (CVPR)*, pages 267–272.

[Ribeiro-Neto e Muntz, 1996] Ribeiro-Neto, B. and Muntz, R. R. (1996). A belief network model for IR. In *ACM Conference on Research and Development in Information Retrieval - SIGIR96*, pages 253–260.

[Ribeiro-Neto et al., 2000] Ribeiro-Neto, B., Silva, I., and Muntz, R. (2000). Bayesian Network Models for IR. In F. Crestani and G. Pasi, editor, *Soft Computing in Information Retrieval Techniques and Applications*, pages 1259–291. Springer Verlag.

[Rosch et al., 1976] Rosch, E., Mervis, C. B., Gray, W., Johnson, D., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychol.*

[Ross, 1998] Ross, S. (1998). *A First Course in Probability.* Prentica-Hall.

[Rowley et al., 1998] Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Transaction on Pattern Analysis and Machine Inteligence*, 20(1):23–38.

[Rubner et al., 1998] Rubner, Y., Tomasi, C., and Guibas, L. J. (1998). A Metric for Distributions with Applications to Image Databases. In *IEEE International Conference on Computer Vision (ICCV'98)*, pages 207–214.

[Russell, 1969] Russell, E. J. (1969). Extension of Dantzig's Algorithm to Finding an Initial Near Optimal-Basis for Transportation Problem. *Operat Res*, 17:187—191.

[Russell e Norvig, 1995] Russell, S. J. and Norvig, P. (1995). *Artificial Inteligence: a moderm approach.* Englewood Cliffss, Prantice-Hall.

[Salton e Buckley, 1990] Salton, G. and Buckley, C. (1990). Improving retrievalperformance by relevance feedback. *Journal of the American Society for Information Science.*, 41:288–297.

[Salton e McGill, 1997] Salton, G. and McGill, M. (1997). *Introduction to Moderm Information Retrieval.* McGraw-Hill Book Co., New York.

[Salton et al., 1997] Salton, G., Wong, A., and Yang, C. S. (1997). A vector spacew model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

[Schiele e Crowley, 1996] Schiele, B. and Crowley, J. L. (1996). Object recognition using multidimensional receptive field histograms. In *Proceedings of Fourth European Conference on Computer Vision*, pages 610–619, Cambridge, UK.

[Schiele e Crowley, 1997] Schiele, B. and Crowley, J. L. (1997). The concept of visual classes for object classification. In *Proceedings of SCIA'97, Thenth Scandinavian Conference on Image Analysis*, pages 43–50, Lappeenranta, Finland.

[Schmidt, 1907] Schmidt, E. (1907). Zur Theorie der linearen und nichttlinearen Integralgleichungen. I Tiel. Entwicklung willkurlichen Funktionen nacht System vorgeschriebener. *Mathematische Annalen*, 63.

[Schneiderman e Kanade, 1998] Schneiderman, H. and Kanade, T. (1998). Probabilistic modeling oflocal appearance and spatial relationships for object recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 98)*, pages 45–51, Santa Barbara, CA.

[Sclaroff et al., 1997] Sclaroff, S., Taycher, L., and Cascia, M. L. (1997). ImageRecover: Acontent-based image browser for the world wide web. In *Proc. of IEEE int. Workshop on Content-Based Access of Image and Video Libraries*.

[Shen e Wong, 1983] Shen, H. C. and Wong, A. K. C. (1983). Generalized texture representation and metric. In *Computer Vision Graphics Image Process*, volume 23, pages 187–206.

[Smith, 1998] Smith, J. R. (1998). Image retrieval evaluation. In *IEEE Workshop on Content-based Access of Images and Video Libraries (CBAIVL'98)*, pages 112–113.

[Smith e Chang, 1996a] Smith, J. R. and Chang, S. F. (1996a). Integrated Spatial and Feature Image Query. *ACM Multimedia*.

[Smith e Chang, 1996b] Smith, J. R. and Chang, S. F. (1996b). VisualSEEk: a fully automated content-based query system. In *In: ACM Multimedia 96*, pages 87–98, Boston, M.A.

[Srihari, 1995] Srihari, R. K. (1995). Automatic indexing content-based retrieval of captioned images. *IEEE Computer*, 28(9):49–56.

[Szummer e Picard, 1998] Szummer, M. and Picard, R. (1998). Indoor-outdoor image classification. In *IEEE International Workshop on Content-based Acess of Image and Video Databases (CAIVD98)*, pages 42–51, Bombay, India.

[Ullman, 1996] Ullman, S. (1996). *High-level Vision, pg. 3.* Massachusetts Institute of Technology.

[Vailaya et al., 1998] Vailaya, A., Jain, A., and Zhang, H. J. (1998). On image classification: city images vs landscapes. *Pattern Recognition*, 31(12):1921—1936.

[Vailaya e Jain, 1999] Vailaya, A. and Jain, A. K. (1999). Incremental learning for Bayesian classification of images. In *IEEE International Conference on Image Processing (ICIP'99)*, Kobe, Japan.

[Vailaya e Jain, 2000] Vailaya, A. and Jain, A. K. (2000). Detecting sky and vegetation on outdoor images. In *Storage and Retrieval for Media Databases*, pages 411–420. Proc. SPIE 3972.

[Vecera e Farah, 1997] Vecera, S. P. and Farah, M. (1997). Is visual image segmentation a bottom-up or a interactive process? *Perception & Psychophsycs*, 59:1280–1296.

[Werman e Peleg, 1984] Werman, M. and Peleg, S. (1984). Multiresolution Texture Signature Using Min-Max Operators. In *7th International Conference on Pattern Recognition*, pages 97–99.

[Werman et al., 1985] Werman, M., Peleg, S., and Rosenfeld, A. (1985). A Distance Metric for Multidimensional Histograms. *Computer Vision Graphics Image Process*, 32(3):328–336.

[Zhou e Huang, 2001] Zhou, X. S. and Huang, T. S. (2001). Edge-based structural features for content-based image retrieval. *Patterns Recognition Letters*, 22:457–468.