

**EXPLORANDO CATEGORIAS VISUAIS PARA
BUSCA DE PRODUTOS EM SITES DE
COMÉRCIO ELETRÔNICO**

RAFAEL CÂMARA ZACARIAS

**EXPLORANDO CATEGORIAS VISUAIS PARA
BUSCA DE PRODUTOS EM SITES DE
COMÉRCIO ELETRÔNICO**

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

ORIENTADOR: JOÃO MARCOS B. CAVALCANTI

Manaus

Março de 2013

Agradecimentos

Agradeço aos prótons por serem tão positivos, aos nêutrons pela sua neutralidade e aos elétrons pela sua carga.

“The best way to predict the future

is to invent it.”

(Jhon Kay)

Resumo

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incident ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

Palavras-chave: Visão Computacional, Redes, Sabotagens.

Abstract

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incident ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur?

Keywords: Computer Vision, Networks, Sabotage.

Lista de Figuras

1.1	Exemplo de consultas com o mesmo objetivo, na forma textual e visual	2
1.2	Exemplo de consulta: "camisa xadrez" no <i>Google Product Search</i>	3
1.3	Exemplo de itens da categoria "camisa/feminina/azul" em <i>Dafiti.com.br</i>	4
1.4	Exemplo de itens da categoria "camisas/masculinas" em <i>kanui.com.br</i>	5
1.5	Exemplo de imagens que pertenceriam à categorias visuais distintas	6
1.6	Matriz representando categorias visuais baseadas em textura	6
1.7	Taxonomia visual utilizada nos experimentos	7
2.1	Exemplos de itens com a palavra-chave "floral"	9
2.2	Comparação do método baseado em imagens(iLike) e palavras-chave com o método puramente textual(Baseline)	10
2.3	Grafo de similaridade gerado com os 1000 primeiros resultados da busca por "Mona-Lisa"	11
2.4	Tabela de médias de imagens irrelevantes por consulta	11
2.5	Resultado comparativo de ANMRR para os descritores de textura com diferentes dimensões	12
2.6	Resultado de precisão-revocação comparativo entre Histograma de Cor, Descritor baseado em cantos e CSS	13
2.7	Resultado de micro-acurácia comparando diferentes configurações de matching	13
3.1	Cubo representando o espaço de cor RGB	16
3.2	Cilindro representando o espaço de cor HSV	17
3.3	Exemplo de redução da saturação de um pixel	17
3.4	Representação da distribuição de cores no espaço YIQ	18
3.5	Exemplo de Taxonomia Visual para outro domínio	19
3.6	Exemplo de instâncias de diferentes categorias, com características visuais comuns dentro de cada categoria	20

3.7	Exemplo de imagem e respectivo Histograma de Cor RGB	21
3.8	Função de distância L1	21
3.9	Função de similaridade de Tanimoto	22
3.10	Fluxo da geração de características através do descritor CEDD	23
3.11	Definição de subimagem e Bloco de Imagem	24
3.12	5 tipos de arestas consideradas no descritor EHD	24
3.13	Vetor de suporte dividindo espaço bi-dimensional	25
3.14	Cálculo da macro-precisão	26
3.15	Cálculo de precisão	26
3.16	Cálculo do MAP - <i>Mean Average Precision</i>	27
4.1	Esquema simplificado do método VisCat	30
4.2	Exemplo de uma taxonomia com diferentes padrões visuais	30
4.3	Esquema do método VisCat	33
5.1	Exemplo de tela do sistema de avaliação	36
5.2	Taxonomia da base de 5.000 camisas masculinas	37
5.3	Macro-Precisão distribuída entre classes	38
5.4	Curva macro-precisão X instâncias de teste por classe	38
5.5	Precisão em 10 para cada consulta com a base de camisas masculinas	39
5.6	Média da precisão em 10, 20 e 30 na base de produtos de camisas masculinas	40
5.7	Precisão em 10 para cada consulta com a base de produtos de vestuário em geral	41
5.8	Média da precisão em 10, 20 e 30 na base de produtos de vestuário em geral	41
5.9	Exemplos de consulta e respostas na base de Roupas Masculinas	43
5.10	Exemplos de consulta e respostas na base de Vestuário em Geral	44

Lista de Tabelas

Sumário

Agradecimentos	v
Resumo	ix
Abstract	xi
Lista de Figuras	xiii
Lista de Tabelas	xv
1 Introdução	1
2 Trabalhos relacionados	9
3 Conceitos básicos	15
3.1 Recuperação de imagem baseada em conteúdo	15
3.1.1 Imagem digital	15
3.2 Categorias Visuais	19
3.3 Descritores de imagem	20
3.3.1 <i>Compact Edge Directivity Descriptor</i> - CEDD	22
3.3.2 <i>Edge Histogram Descriptor</i> - EHD (MPEG-7)	23
3.4 Classificação de imagens com Máquinas de Vetor Suporte (SVM's)	25
3.5 Métricas de avaliação	25
3.5.1 Macro-Precisão	26
3.5.2 Precisão em N	26
3.5.3 MAP- <i>Mean Average Precision</i>	27
4 O Método	29
4.1 Explorando Categorias Visuais- O método VisCat	29
4.1.1 Construção da Taxonomia Visual	30

4.1.2	Busca em escopo reduzido	32
5	Experimentos	35
5.1	Configuração do Experimento	35
5.2	Classificação da base de imagens	36
5.3	Busca visual de imagens	39
5.3.1	Coleção de Camisas Masculinas - 5.000 imagens	39
5.3.2	Coleção de Vestuário geral - 23.000 imagens	40
5.4	Comentários	41
6	Conclusões e trabalhos futuros	45
	Referências Bibliográficas	47
	Anexo A Lista de sites coletados para montagens das bases de imagens	51

Capítulo 1

Introdução

As inovações das tecnologias multimídia aumentaram a disponibilidade de imagens digitais ao longo dos últimos anos. Os sistemas de comércio eletrônico também cresceram e com isso, criou-se uma demanda de novas soluções baseadas em imagens para estes ambientes (Chen et al. [2010]Jing & Baluja [2008]), em particular para busca, classificação, catalogação e outras necessidades baseadas em componentes visuais. Desta forma, começou-se a empregar técnicas de CBIR (*Content-Based Image Retrieval*) para busca de produtos em comércio eletrônico Jing & Baluja [2008].

No âmbito das diferentes categorias de comércio eletrônico, a categoria de peças de vestuário (roupas, calçados e acessórios) tem grande apelo visual, o que afeta diretamente a decisão de compra por parte do usuário, diferentemente de produtos eletro-eletrônicos e seus derivados, por exemplo. Ao expor/recuperar estes dados (imagens de produtos de vestuário) de maneira mais apropriada ao interesse do usuário, aumentam as chances de uma compra ser realizada.

Em linhas gerais, algoritmos de classificação ou agrupamento podem ser usados para melhorar resultados de uma busca visual(Iwayama & Tokunaga [1995]). Em um sistema de busca simples, sem a abordagem de classificação, cada consulta é confrontada com toda a base. Ao se utilizar a classificação, cada consulta pode ser confrontada apenas com subconjuntos da base, sem prejudicar a eficácia.

Além de se reduzir os cálculos necessários para as comparações entre a imagem de consulta e as imagens da base, se a categoria em que a consulta foi inferida estiver correta, há uma maior chance de que mais itens relevantes sejam recuperados no topo da resposta. Analogamente, o uso de classificação de imagens de produtos em categorias visuais pode melhorar显著mente o resultado de uma busca visual, e como a busca é uma das principais ferramentas para se encontrar produtos de interesse, pode-se inferir que uma classificação com alta precisão pode ser uma forte ferramenta para

sistemas deste contexto.

Com a crescente demanda por soluções baseadas em CBIR nos últimos anos, a busca visual torna-se uma ferramenta promissora para novos contextos. Dentro das abordagens existentes para sistemas de busca, temos a busca vertical, que consiste em recuperar objetos dentro de um contexto específico. Um dos contextos em que a busca vertical pode ser útil, é o de comércio eletrônico. No caso deste trabalho⁶, o método será aplicado em uma base de produtos de vestuário, mas a aplicação pode ser feita em outros domínios de busca visual, exigindo o cuidado de fazer as adaptações necessárias.

A principal razão de se utilizar busca visual é que uma imagem pode representar melhor a consulta do usuário do que com informações textuais. Na Figura 1.1, temos um exemplo, onde duas consultas com o mesmo objetivo são descritas de forma visual e textual. Percebe-se que o esforço de um usuário para digitar as características visuais (à esquerda) seria muito maior do que se um exemplo visual fosse dado como consulta (à direita).



Figura 1.1. Exemplo de consultas com o mesmo objetivo, na forma textual e visual

Como exemplo de sistemas que usam classificação de produtos de vestuário, temos várias lojas de comércio eletrônico como Dafiti¹, Kanui², Netshoes³ e Lojas Leader⁴. Além destas lojas, há sistemas que reunem informações sobre produtos em diversas fontes, como o *Google Product Search*⁵, que além de exibir produtos de vestuário relacionados à consulta textual, disponibiliza a classificação de produtos da resposta através de formas visuais. Na Figura 1.2 podemos observar uma tela deste sistema. É possível

¹www.dafiti.com.br

²www.kanui.com.br

³www.netshoes.com.br

⁴www.lojasleader.com.br

⁵www.google.com/shopping

ver as imagens dadas como respostas para a consulta textual, bem como a opção de filtragem dos resultados por categorias de camisas à esquerda, como mangas curtas e longas.

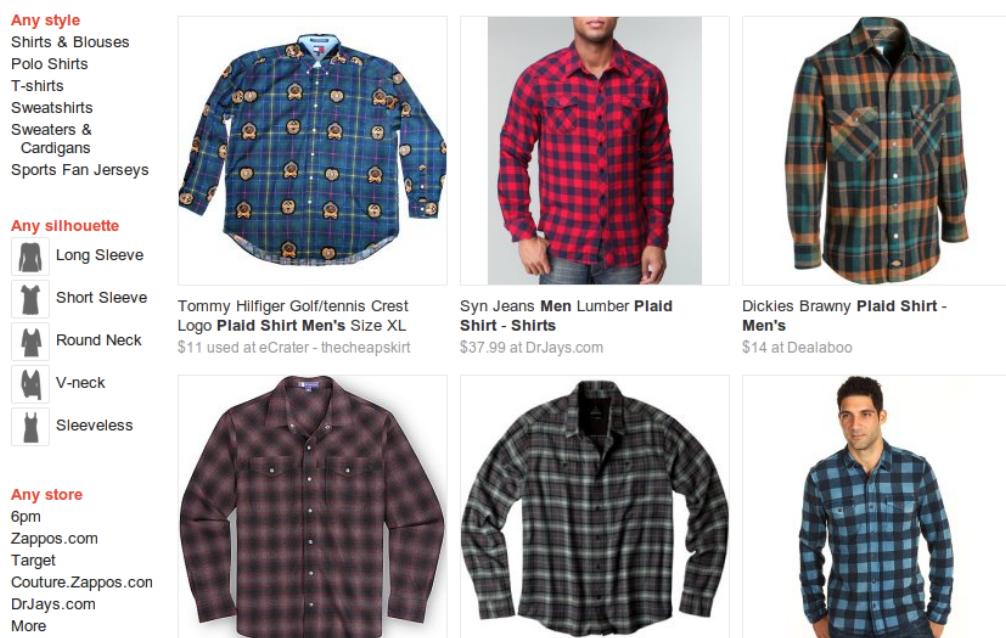


Figura 1.2. Exemplo de consulta: "camisa xadrez" no *Google Product Search*

Outro sistema que classifica as imagens de produtos por informações visuais é a loja online Dafiti.com.br. Neste ambiente, as peças de vestuário podem ser classificadas por informações visuais de cor. Na Figura 1.3 é possível ver uma categoria de camisas femininas. À esquerda, é possível filtrar os itens por cor predominante.

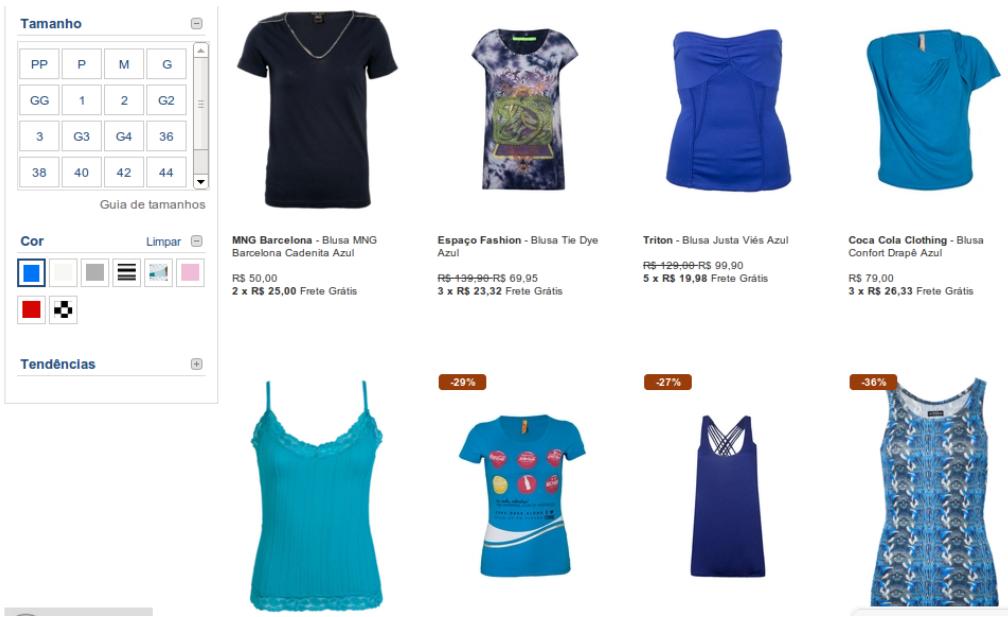


Figura 1.3. Exemplo de itens da categoria "camisa/feminina/azul" em *Dafiti.com.br*

Os sites de comércio eletrônico utilizam categorias para facilitar a catalogação e a visualização de seus produtos. Conforme as Figuras 1.2 e 1.3, podemos verificar que cada site tem uma classificação com sua própria taxonomia, mas não há um padrão que seja utilizado por todos estes sistemas. Por exemplo, em itens de vestuário na loja Dafiti (Figura 1.3), a taxonomia da categoria de camisa se divide por cor predominante. Já na loja Kanui (Figura 1.4), é possível visualizar os produtos baseado no tipo de manga ou no tecido que compõe o item.

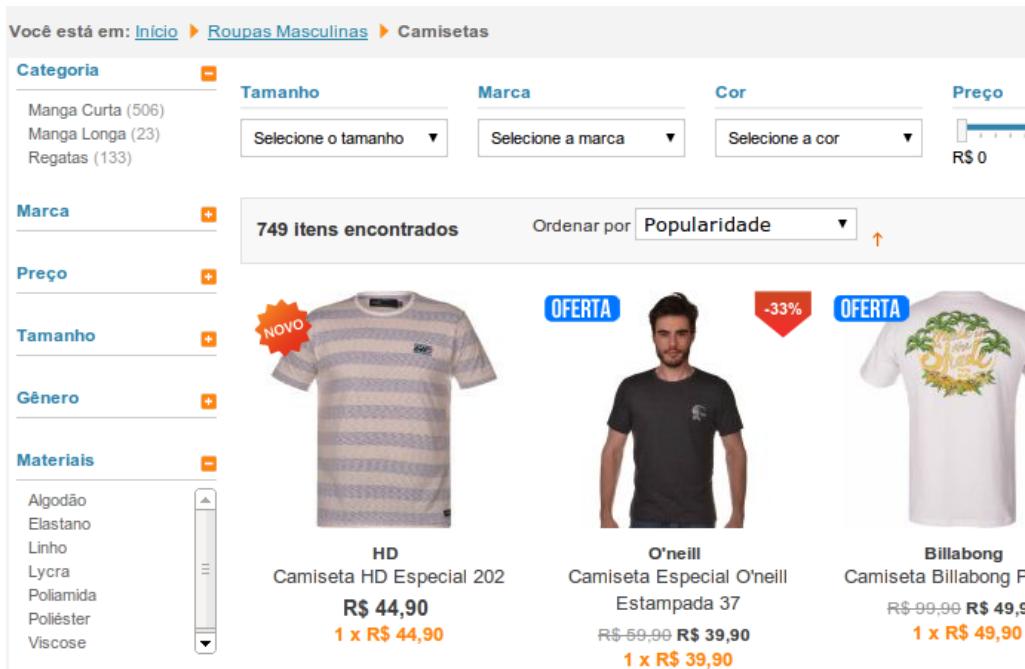


Figura 1.4. Exemplo de itens da categoria "camisas/masculinas" em kanui.com.br

Neste trabalho, o método proposto *Visual Categories* - VisCat - é uma abordagem para busca de imagens baseada em conteúdo, que usa um esquema simples de combinação de classificação e busca, com descritores já conhecidos da literatura, mas que são eficientes em coleções que apresentem informações de categorias visuais bem definidas.

Quando se trata de categorias visuais, pode-se utilizar algum padrão visual para determinar a distribuição de instâncias nestas categorias. O padrão visual pode ser obtido através de informações baseadas no conteúdo de uma imagem, como cor, forma ou textura, conforme o contexto da aplicação. Como este trabalho tem o foco em comércio eletrônico de produtos de vestuário, a informação de textura foi escolhida como parâmetro para definição das classes. Na Figura 1.5 é possível ver três exemplos de produtos, que poderiam compor três categorias visuais: listradas, xadrez e lisas, respectivamente.



Figura 1.5. Exemplo de imagens que pertenceriam à categorias visuais distintas

A abordagem deste trabalho incorpora informação de cor e textura nas suas diferentes etapas. Na classificação, são utilizadas as informações de textura para descrever imagens, gerar seus vetores de características e classificá-las, utilizando *Support-Vector Machines (SVM)* Crammer & Singer [2002]. As informações de cor não são utilizadas nesta etapa, pois para a taxonomia que será proposta, não será interessante que imagens com um mesmo padrão visual de textura (xadrez, por exemplo) e cores diferentes sejam classificadas em classes distintas. Na Figura 1.6 é possível ver imagens de mesmas cores e padrão de textura diferentes, e as classes pretendidas são as linhas da matriz. Na fase de busca, dada a classificação de uma imagem de consulta, é realizada a busca diretamente na classe prevista para a consulta, porém utilizando adicionalmente a informação de cor, para recuperar imagens com cores similares.

	Cor Vermelha	Cor Preta	Cor Verde
Camisa Xadrez			
Camisa sem Estampa			
Camisa Listrada			

Figura 1.6. Matriz representando categorias visuais baseadas em textura

Os experimentos foram divididos em duas etapas: (i) avaliamos resultados experimentais da classificação das imagens em categorias visuais. Através da macro-precisão foi possível mensurar a eficiência do método de classificação. Foram utilizadas 5000 imagens de roupas masculinas coletadas de 20 diferentes sites de comércio eletrônico. Os resultados obtidos mostraram que com um simples descritor de textura foi possível alcançar um alto grau de macro-precisão (89%) na coleção proposta. (ii) Foi avaliado a precisão em uma busca visual com 50 consultas nesta base de 5.000 imagens e também em uma base heterogênea, 23.0000 imagens de vestuário em geral, onde é possível ver o ganho do VisCat na base homogênea, em comparação com outros descritores utilizados como baselines. Na Figura 1.7 é possível ver a taxonomia de 5 categorias visuais utilizada neste experimento, com uma imagem de exemplo para cada classe.



Figura 1.7. Taxonomia visual utilizada nos experimentos

As contribuições deste trabalho são duas. (i) demonstrar como descritores *CBIR* pouco complexos são suficientes para serem empregados na tarefa de classificação de imagens de produtos de vestuário com um alto grau de precisão; e (ii) mostrar que o uso desta classificação melhora significativamente os resultados de uma busca visual por um produto.

Esta dissertação está organizada em 6 capítulos. No capítulo 2 serão apresentados os trabalhos relacionados. Em seguida, no capítulo 3 descreveremos conceitos básicos necessários para a compreensão deste trabalho. No capítulo 4 será descrita a proposta do método de busca visual combinada com a classificação em categorias visuais. O capítulo 5 descreve os resultados dos experimentos realizados para avaliar o desempenho do método proposto, em uma base com imagens extraídas de lojas de comércio eletrônico. Por último, no capítulo 6 apresentamos nossas conclusões, as contribuições do trabalho desenvolvido e são apresentadas sugestões para trabalhos futuros.

Capítulo 2

Trabalhos relacionados

A área de busca visual de imagens é um segmento de pesquisa recente. Em Chen et al. [2010], temos o *iLike*, que é um exemplo de sistema de busca visual vertical, cujo o contexto é o ramo de vestuário. Neste trabalho, a tecnica consistia em coletar paginas Web, e a partir dos textos destas paginas, eram associadas palavras-chave as imagens desta pagina. Em seguida, foram computadas características visuais a partir das propriedades das imagens de uma mesma palavra-chave, ou seja, foram mapeadas em espaços visuais. Na Figura 2.1 é possível observar um exemplo de mapeamento de imagens em palavras-chave. As características visuais foram usadas para adicionar informação na computação do ranking. Desta forma, cada consulta feita textualmente ao sistema foi respondida visualmente com base nas palavras-chave presentes na consulta. A principal vantagem descrita foi a recuperação de itens relevantes para a consulta, mesmo que estes itens não tivessem palavras-chave da consulta textual.



Figura 2.1. Exemplos de itens com a palavra-chave “floral”

Foi utilizada a métrica de precisão-revocação, comparando este sistema com uma busca puramente textual. Foram realizadas 50 consultas, e cada consulta era expandida visualmente e textualmente para recuperação de 30 itens, em uma base de aproximadamente 20.000 itens. Cada item foi avaliado como relevante ou não, permitindo então o cálculo do valor de precisão-revocação. Na Figura 2.2 podemos ver o ganho do método proposto sobre a busca textual.

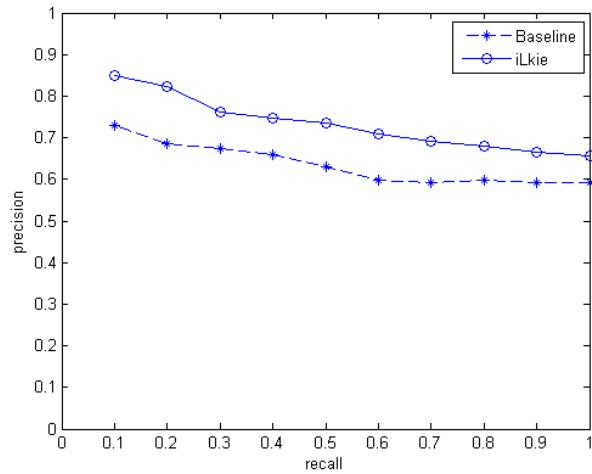


Figura 2.2. Comparação do método baseado em imagens(iLike) e palavras-chave com o método puramente textual(Baseline)

Em Jing & Baluja [2008], temos um método para melhorar os resultados visuais de produtos para uma consulta textual. A técnica, chamada de PageRank, consiste em basicamente em calcular as similaridades entre 1000 imagens do topo do ranking da busca recuperadas pelo Google. As imagens com maior grau de similaridade serão movidas para as primeiras posições do topo do ranking. Uma representação através de um grafo pode ser vista na Figura 2.3.



Figura 2.3. Grafo de similaridade gerado com os 1000 primeiros resultados da busca por "Mona-Lisa"

Neste caso, para a consulta "Mona-Lisa" foram selecionados 1000 imagens do topo do ranking. Após os cálculos de similaridade, temos as duas maiores imagens (centro) com maior pontuação no ranking final. Foi utilizado o SIFT para o cálculo de similaridades, porém o autor ressalta que outros métodos/descritores poderiam ter sido utilizados.

Para avaliar a proposta, foi calculada a média de imagens irrelevantes recuperadas por consulta, no resultados do topo de 10, 5 e 3 respostas, que pode ser visto na Figura 2.5. Assim como no trabalho relacionado anterior, foram utilizadas 50 consultas para esta avaliação, com 150 voluntários.

	Image Rank	Google
<i>Among top 10 results</i>	0.47	2.82
<i>Among top 5 results</i>	0.30	1.31
<i>Among top 3 results</i>	0.20	0.81

Figura 2.4. Tabela de médias de imagens irrelevantes por consulta

O autor ainda ressalta que a esta técnica pode ser utilizada em aplicações que consistam na seleção de um conjunto pequeno de imagens em um grande conjunto de candidatos.

Em Kejia et al. [2011], temos o uso de descritores de imagens baseados forma, para recuperação de imagens de produtos em geral. Estes descritores foram utilizadas baseando-se na ideia de que é possível encontrar o contorno do principal objeto das imagens, tanto das consultas, quanto da base. Sendo assim, torna-se possível inferir que objetos com formas parecidas são da mesma classe. O artigo é focado na comparação de 3 descritores de forma baseados em momentos: *Zernike*, *Jacobi-Fourier* e *Radial-Harmonic-Fourier*.

Os experimentos foram feitos com um conjunto de 16 categorias, em que cada categoria possuía 100 imagens na base e 20 consultas cada. O valor de ANMRR foi calculado para a avaliação dos descritores, conforme a Figura 2.5. Concluiu-se que o descritor *Radial-Harmonic-Fourier* tem desempenho superior aos outros dois métodos.

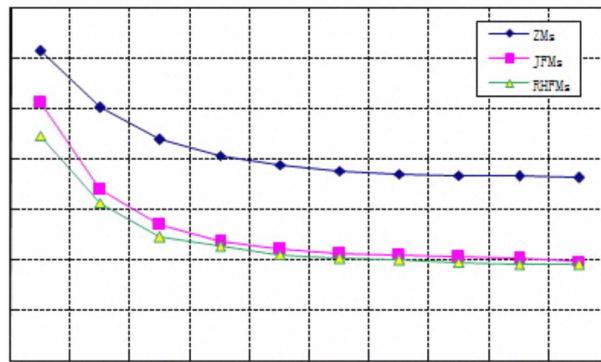


Figura 2.5. Resultado comparativo de ANMRR para os descritores de textura com diferentes dimensões

Em Tseng et al. [2009], utilizou-se também as evidências de forma, baseadas em contorno e seleção de pontos de interesse nas imagens. Neste caso, o problema consistia em realizar a busca visual por peças de vestuário. Pelo método de seleção de pontos de interesse, verificou-se que imagens de vestuário podem conter muitos ruídos como oclusão, sombra, iluminação, dobras das peças, e outras deformidades visuais que podem prejudicar a extração de características dos descritores de imagem.

Em Ouyang [2009], é proposta uma técnica para busca visual em uma base de roupas de vestuário. O autor utiliza informação textual de *tags* atribuídas a imagens, para construir uma árvore semântica para categorizar os produtos, chamada pelo autor de *Clothes Semantic Struture* (CSS). Com um descritor baseado em pontos de interesse (SIFT), é possível adicionar informação visual na árvore semantica, facilitando a categorização das imagens da base e da consulta. Foi utilizada uma base com 5.000 imagens, distribuídas em 9 categorias. O número de consultas não consta no artigo. Na Figura 2.6, pode-se ver a tabela comparativa entre o método proposto (visual-textual)

e outros dois descritores visuais.

<i>Appraisal</i>	<i>Color histogra m</i>	<i>Corner s-based</i>	<i>Proposed method</i>
Precision	76%	83%	89%
Recall	69%	74%	81%

Figura 2.6. Resultado de precisão-revocação comparativo entre Histograma de Cor, Descritor baseado em cantos e CSS

Em Xie et al. [2008], também temos o uso de um descritor baseado em pontos de interesse (SIFT), porém para categorização de produtos de comércio eletrônico em geral. A base possuía 12.000 imagens, distribuídas uniformemente em 100 categorias. Das 120 imagens de cada categoria, 100 eram utilizadas como base de treino e 20 eram utilizadas como teste. Desta forma, 2.000 itens deveriam ser corretamente classificados, a partir de um aprendizado com 10.000 itens. Na Figura 2.7, é possível ver a comparação dos valores de micro-acurácia obtidos em diferentes configurações de matching. As legendas N-N/N-1/AR correspondem ao método de filtragem de vizinhos próximos.

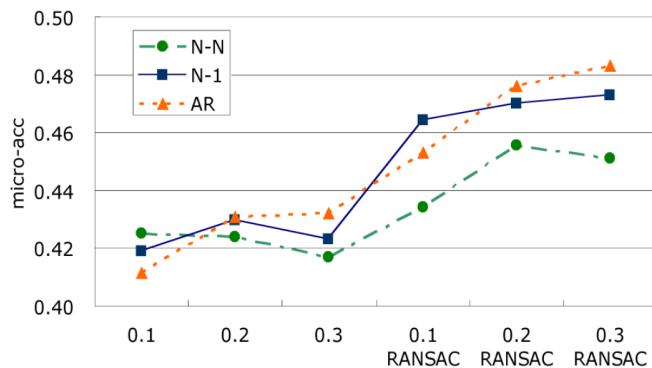


Figura 2.7. Resultado de micro-acurácia comparando diferentes configurações de matching

Dentre os trabalhos relacionados apresentados, o último é, essencialmente, o mais parecido com a nossa proposta: classificação automática de produtos. A grande diferença consiste na aplicação (produtos em geral x roupas) e nas técnicas utilizadas para solucionar o problema de classificação: descritores (SIFT X CEDD+EHD) e algoritmos de aprendizagem de máquina empregados (RNA's x SVM's).

Capítulo 3

Conceitos básicos

Neste capítulo serão definidos os conceitos necessários para o entendimento do trabalho. Primeiramente algumas definições de CBIR serão explicadas, como imagem digital, espaços de cor e descritores. Em seguida, uma breve explicação sobre máquinas de vetor suporte. Por último, serão apresentadas as métricas utilizadas para avaliar a eficácia dos métodos empregados neste trabalho.

3.1 Recuperação de imagem baseada em conteúdo

Nesta seção apresentaremos os conceitos de Recuperação de Imagem Baseada em conteúdo utilizados para o desenvolvimento deste trabalho. Primeiramente o conceito básico de imagem digital será explicado, em seguida serão apresentados conceitos relacionados, como espaços de cor e descritores de imagem.

3.1.1 Imagem digital

No contexto de CBIR, consideramos a representação primária de uma imagem digital na forma de matriz de pixels, onde cada pixel tem um valor que representa uma cor, dentro de uma estrutura chamada espaço de cor. Neste trabalho, usamos os espaços de cor o HSV(*Hue Saturation Value*), o YIQ (*Luminance In-phase Quadrature*), e o RGB (Red, Green, Blue), que serão descritos a seguir.

Espaço de cor RGB - *Red, Green, Blue*(Vermelho, Verde, Azul)

No espaço de cor RGB, cada pixel será composto por três valores, um referente a tonalidade de vermelho, um a tonalidade de verde e o último, referente a tonalidade de azul. Pode-se inferir que para se representar uma imagem digital I através do esquema

de cor RGB, utiliza-se uma matriz de pixels, mas ao invés de termos valor único para cada um destes pixels (posição da matriz), teremos três valores, um para cada banda de cor. Destes três componentes de um pixel, seus valores estarão entre 0 e 255, onde zero significa a ausência total desta cor, e 255 a tonalidade mais forte desta cor. Na Figura 3.1, o espaço de cor RGB está representado por um cubo de três dimensões, onde os valores R, G, e B são respectivamente as tonalidades de Vermelho, Verde e Azul. O sentido das setas na figura indicam o aumento do valor em cada eixo.

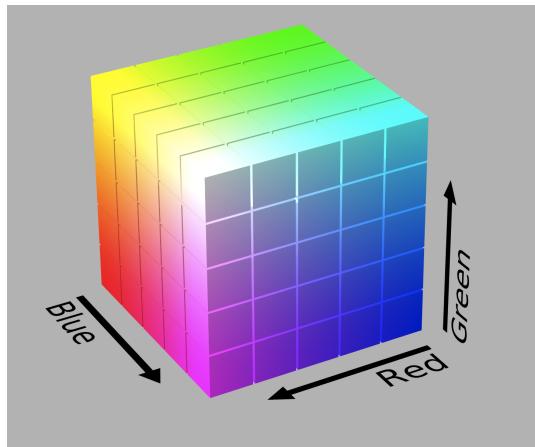


Figura 3.1. Cubo representando o espaço de cor RGB

Da figura anterior, podemos notar que a distribuição de cores pode ser observada de forma tridimensional, e cada ponto do cubo de cores é um mapeamento de uma cor formada pelos 3 componentes. Desta forma, os valores de RGB de certas cores podem ser inferidos: o tom de azul mais forte será $\text{RGB}=(0,0,255)$; o de verde será $\text{RGB}=(0,255,0)$ e o de vermelho será $\text{RGB}=(255,0,0)$. Outro aspecto pode ser observado a partir do cubo de cor RGB: todos os pixels que forem situados em uma diagonal que se inicie da origem dos três componentes, será um tom de cinza (valores de R, G e B iguais). Quanto maior forem os valores de RGB de um pixel em tom de cinza, mais próximo será da cor branca.

Espaço de cor HSV- *Hue Saturation Value (Tonalidade-Saturação-Valor)*

O espaço de cor HSV é fundamentalmente diferente do espaço de cor RGB, uma vez que é separada a Intensidade(luminância) da informação de cor (cromaticidade). Nos dois eixos de cromaticidade, a diferença no valor de Tonalidade (*Hue*) de um pixel é visualmente mais significativo comparado com o que ocorre quando há a variação de Saturação (Sural et al. [2002]). Desta forma, o valor de Tonalidade se torna semânticamente mais descriptivo, ou seja, uma variação dos três valores que compõem um pixel

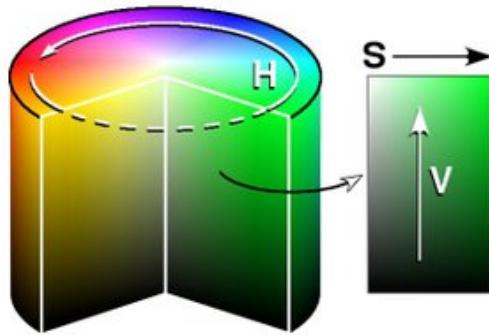


Figura 3.2. Cilindro representando o espaço de cor HSV

no espaço HSV, a variação em Saturação (*Saturation*) e em Valor (*Value*) são menos importantes. Na Figura 3.2, o espaço de cor HSV está representado por um cilindro de três dimensões, onde H representa Tonalidade, S representa Saturação e V representa Valor. O componente H refere-se a dimensão angular, começando com vermelho em 0° , passando pelo verde em 120° , em seguida pelo azul em 240° , e retornando ao vermelho em 360° . Já os componentes V e S constituem no deslocamento porcentual nestes dois eixos para cada ângulo.

A utilização do espaço de cor HSV se iniciou pela necessidade de se representar variação perceptual de cores com coordenadas mais intuitivas. Para exemplificar esta variação, na Figura 3.3 temos a transformação de um tom de laranja saturado (fortemente colorido) para um tom de laranja com menos saturação - no espaço HSV, apenas o valor de S varia nesta transformação. Por outro lado, para representar a mesma transformação no espaço de cor RGB, o valor de R deveria ser reduzido em 31 unidades, o G aumentado em 24, e o B aumentado em 59.

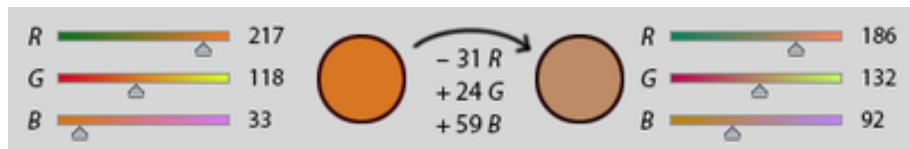


Figura 3.3. Exemplo de redução da saturação de um pixel

Como a distribuição R,G,B não representa cores perceptualmente semelhantes com valores próximos, o espaço de cor HSV torna-se adequado para descritores baseados em cor, onde podem ser utilizados sistemas *fuzzy* para normalizar os valores de cada pixel, e assim preservar a proximidade de pixels semânticamente parecidos, e acentuar a diferença entre pixels com cores distantes. Quando existe variação na iluminação,

por exemplo, um pixel HSV possuirá valores de intensidade (*Value*) diferentes, mas os componentes de cor (*Hue* e *Saturation*) serão próximos.

Espaço de cor YIQ- *Luminance In-Phase Quadrature*(Luminância-Fase-Quadratura)

O espaço de cor YIQ foi o sistema adotado pelo *National Television System Committee (NTSC)* para transmissão do sinal analógico de televisão. Este modelo foi escolhido partindo do pressuposto de certas características da visão humana, maximizando a utilização de uma banda fixa de transmissão: a visão humana é mais sensível a mudanças na luminância do que em mudanças na Tonalidade ou na Saturação, e então uma banda mais larga do sinal deveria ser dedicada à informação de luminância, e não na informação de cor. O eixo Y pode ser entendido como luminância, e I e Q trazem informação de cromaticidade. Nos modelos de televisores preto-e-branco, apenas o componente Y era utilizado.

Quando imagem está no espaço de cor YIQ, as informações de textura serão facilmente obtidas: ao se descartar os componentes I e Q, ignoram-se as informações de cor, que não seriam úteis para se calcular informações de textura. Estas informações podem ser obtidas exclusivamente do componente Y de uma imagem. Na Figura 3.4, o espaço de cor YIQ está representado por um cubo de três dimensões, onde Y é o valor de Luminância, I é o valor de Fase e Q é o valor de Quadratura.

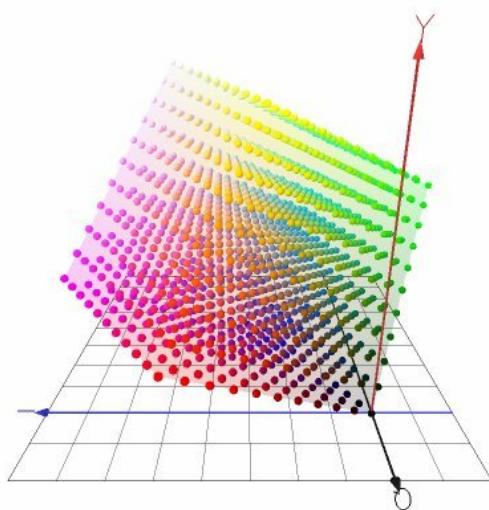


Figura 3.4. Representação da distribuição de cores no espaço YIQ

3.2 Categorias Visuais

Uma categoria visual é um conjunto de imagens que possuem um mesmo padrão visual. O uso de categorias visuais é empregado em diversas abordagens CBIR. Em Fei-Fei et al. [2003], temos um sistema em que categorias visuais foram utilizadas para melhorar uma consulta genérica por imagens. Alguns exemplos de categorias deste trabalho são imagens de garrafas, camelos e carros. Em Fritz & Schiele [2008], temos um sistema em que o objetivo é a classificação não-supervisionada de imagens em categorias visuais. As categorias visuais também eram genéricas, como imagens de motocicletas, pessoas, e ovelhas. Em Vijayanarasimhan & Grauman [2008], também temos um sistema em que o objetivo é a classificação automática. As classes também eram genéricas, como aviões, carros e guitarras. De forma geral, o uso de categorias visuais para busca e classificação é recorrente, e é empregada em contextos genéricos.

As categorias visuais também podem ser baseadas em um tema específico. Neste trabalho o contexto é o comércio eletrônico de roupas de vestuário, então podemos utilizar padrões visuais como "xadrez", "listrado" ou "estampado" para a definição das categorias visuais. Um exemplo de outro contexto seria um sistema de busca de imagens de peixes catalogados, onde cada categoria visual poderia ser baseado em evidências de forma ou cor. Uma possível taxonomia visual para este sistema é apresentada na Figura 3.5, onde na Categoria 1 seriam os peixes com formas arredondadas, na Categoria 2 os peixes médios listrados e na Categoria 3, os peixes compridos.

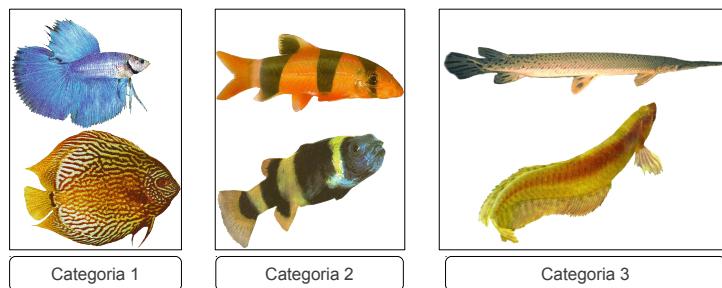


Figura 3.5. Exemplo de Taxonomia Visual para outro domínio

Outro exemplo é mostrado na 3.6, onde a definição da classe se dá pela existência ou não de um item no objeto presente na imagem (fivela) que pode ser descrito visualmente. Com classes e instâncias bem definidas, os vetores de características de imagens de diferentes classes serão suficientemente distintos, o que torna o aprendizado mais preciso.



Figura 3.6. Exemplo de instâncias de diferentes categorias, com características visuais comuns dentro de cada categoria

3.3 Descritores de imagem

Dada a representação matricial de uma imagem, podem-se extrair diversas propriedades visuais, como cor, forma e textura. Na extração de propriedades, um conjunto pré-definido de características representativas do conteúdo de uma imagem deve ser extraído. Este processo é necessário para representar compactamente um grande conjunto de imagens, sem perder qualidade semântica de informações visuais. Para isso, utilizamos vetores de características, que servirão para guardar informações mais relevantes e objetivas do que a simples matriz de pixels RGB.

Um descritor é uma representação de uma imagem através de suas características, sendo composto por um vetor de características \mathbf{v} e uma função de distância \mathbf{d} , usada para calcular um valor de similaridade entre vetores de características. Como os descritores se baseiam nas características visuais de uma imagem, os mais comuns são os de cor, forma e de textura. Para exemplificar um descritor, podemos utilizar um histograma de cor. Na Figura 3.7, temos a demonstração de um Histograma de Cor RGB para uma imagem definida por uma matriz de 6x6 pixels.

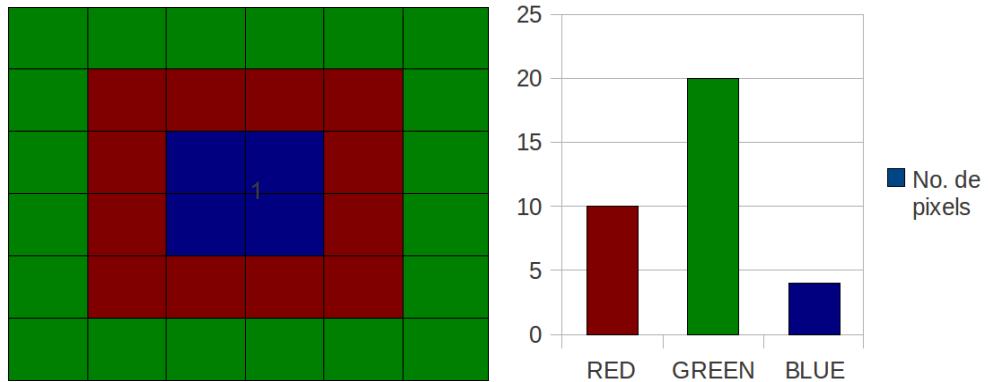


Figura 3.7. Exemplo de imagem e respectivo Histograma de Cor RGB

Conforme a ilustração, podemos observar que um histograma é um vetor com n posições, em que cada posição corresponde a quantidade de pixels daquela cor presente na imagem. O histograma pode ser global ou local: global quando diz respeito à imagem como um todo e local quando diz respeito à uma partição(região) da imagem. De posse do vetor formado pelo histograma da imagem, é necessário escolher uma função de distância **d** ou similaridade para mensurar a diferença ou igualdade entre as imagens representadas pelos histogramas. Existem diversas funções que podem ser utilizadas para este fim. Na Figura 3.8 temos a função de distância L1, onde quanto menor for o valor do resultado, mais próximos são os vetores de entrada. Para utilizar esta função, basta tomar **q** e **p** como componentes dos vetores de características destas imagens. Em nossos experimentos, utilizamos a função similaridade de Tanimoto, onde quanto maior for o resultado, mais próximos serão os vetores de entrada. Esta função pode ser vista na Figura 3.9. Para utilizar esta função, basta tomar **A** e **B** como componentes dos vetores de características destas imagens.

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^n |p_i - q_i|$$

Figura 3.8. Função de distância L1

$$f(A, B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}$$

Figura 3.9. Função de similaridade de Tanimoto

A seguir, descreveremos os descritores de imagens que foram usados neste trabalho.

3.3.1 *Compact Edge Directivity Descriptor* - CEDD

O descritor CEDD Chatzichristofis & Boutalis [2008] é um descritor de características de baixo nível, que incorpora informação de cor e textura em um histograma. Como o tamanho de um histograma CEDD é limitado a 54 bytes por imagem, este descritor possui baixo custo computacional e pode ser utilizado em larga escala, ou seja, pode ser executado em tempo aceitável em grandes bases de imagens.

A unidade associada com a extração de informação de cor é chamada de Unidade de Cor. analogamente, a Unidade de Textura é a unidade associada à informação de textura. O histograma CEDD é constituído por 6 regiões, determinada pela Unidade de Textura. Cada uma dessas 6 regiões é subdividida em 24 subregiões baseadas na Unidade de Cor. O histograma final possui $6 \times 24 = 144$ regiões. Para montar o histograma, primeiramente separa-se a imagem em 1600 Blocos de Imagem. Este número foi proposto para otimizar o compromisso entre a detecção de detalhes de uma imagem e o custo computacional para se obter tais detalhes. Cada Bloco de Imagem alimenta sucessivamente todas as unidades. Se nós definimos que o bin que resulta da Unidade de Textura como T , e o bin que resulta da Unidade de Cor como C , então o Bloco de Imagem é colocado na posição do histograma final $T \times 24 + C$. Na Figura 3.10 temos o fluxo para a geração do vetor de características de um Bloco de Imagem através do CEDD.

Unidade de Textura

Na Unidade de Textura, o Bloco de Imagem é separado em 4 regiões, chamadas de Sub Blocos. O valor de cada Sub Bloco é o valor médio da luminosidade dos pixels que o compõem. Os valores de luminosidade são derivados da transformação através do espaço de cor YIQ (onde Y representa luminância, e I e Q representam crominância). Cada Bloco de Imagem passa por 5 filtros, que servem para classificar as arestas deste bloco com grau 0, 45, 90, 135 ou como aresta não-direcional.

Unidade de Cor

Na Unidade de Cor, cada bloco de Imagem é transferida para o espaço de cor HSV. A média dos valores de H, S e V são calculadas e dadas como entrada em 3 sistemas fuzzy, um para cada componente. A combinação do resultado da saída destes 3 sistemas classificarão o bloco em 27 bins. O processo é repetido para todos os blocos da imagem. Ao fim do processo, o histograma é normalizado no intervalo 0-1. Cada valor do histograma então é quantizado em 3 bits.

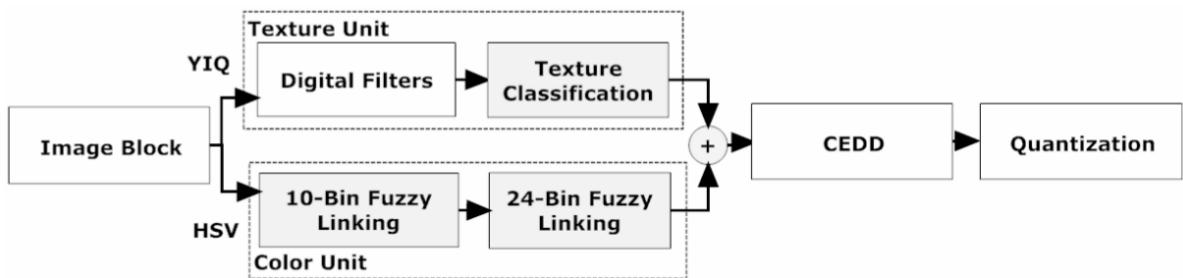


Figura 3.10. Fluxo da geração de características através do descriptor CEDD

3.3.2 *Edge Histogram Descriptor - EHD (MPEG-7)*

O padrão MPEG-7, formalmente denominado "Interface de Descrição de Conteúdo Multimídia", é um padrão para descrever conteúdo de dados multimídia que suportem algum grau de interpretação de significado da informação, que pode ser transferidos para, ou acessados por, um dispositivo ou um código de computador. MPEG-7 não foi destinado para nenhuma aplicação em particular; ao invés disso, os elementos que o MPEG-7 padroniza suporta um grande leque de possíveis aplicações.

O descriptor de histograma de arestas - EHD Park et al. [2000] - representa a distribuição espacial de 5 tipos de arestas (4 direcionais e um não-direcional). Uma vez que arestas tem um papel importante para a percepção de imagens, este método pode recuperar imagens com significância semântica similar, especialmente para imagens naturais com distribuição não-uniforme de arestas. Neste contexto, o desempenho da recuperação de imagem pode ser significativamente melhorada se o descriptor de histograma de aresta for combinado com outros descritores como o histograma de cor. Além disso, os melhores resultados em recuperação de imagens considerando esse descritor sozinho, são obtidos usando histogramas globais e semi-globais, gerados diretamente pelo histograma de arestas assim como o local para o processo de matching.

Partição do Espacial da Imagem para a identificação e localização de arestas

Para localizar a distribuição de arestas em uma certa região de uma imagem, nós dividimos espacialmente a imagem em 16 subimagens (4×4), conforme a Figura 3.11. Então, para cada sub-imagem, geramos um histograma de arestas para representar a distribuição de arestas na sub-imagem. Para definir os tipos diferentes de arestas, cada sub-imagem será dividida em pequenos blocos quadrados chamados de blocos de imagem.

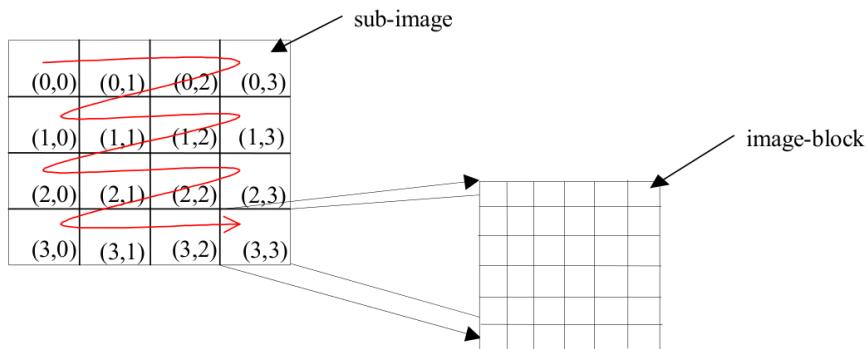


Figura 3.11. Definição de subimagem e Bloco de Imagem

Tipos de Arestas

No histograma do EHD, 5 tipos de arestas são definidos. Há quatro arestas direcionais e uma aresta não direcional. As quatro arestas direcionais incluem vertical, horizontal, e diagonais de 45 graus e 135 graus. Estas arestas direcionais são extraídas dos blocos de imagem. Se o bloco de imagem contém uma aresta arbitrária sem direcionalidade alguma, esta então será classificada como uma aresta não-direcional. Na Figura 3.12 podemos visualizar os diferentes tipos de arestas.

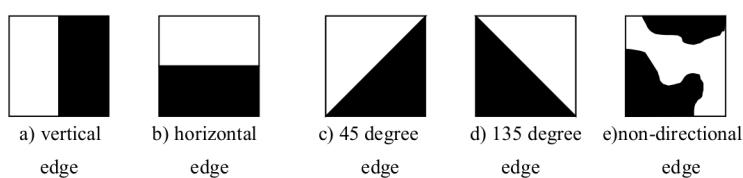


Figura 3.12. 5 tipos de arestas consideradas no descriptor EHD

3.4 Classificação de imagens com Máquinas de Vetor Suporte (SVM's)

Esta é uma técnica de aprendizagem de máquina supervisionada para classificação, onde são necessários exemplos previamente identificados para construir um modelo Cortes & Vapnik [1995]. Neste método, as características dos objetos a serem classificados são transformadas em vetores numéricos, onde cada componente deste vetor corresponde a uma característica. Estes vetores são mapeados em um espaço com alta dimensionalidade através de um mapeamento não-linear. Para que neste novo espaço seja encontrado, o algoritmo calcula um hiperplano ótimo que separe os vetores mapeados em suas diferentes classes. Um exemplo em duas dimensões pode ser visto na Figura 3.13.

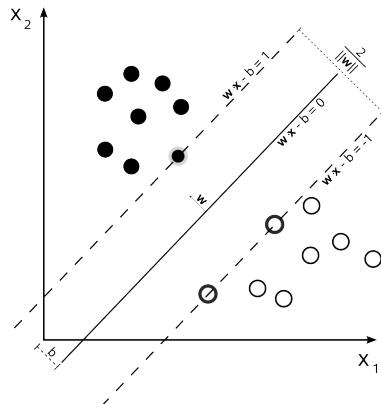


Figura 3.13. Vetor de suporte dividindo espaço bi-dimensional

Em nosso trabalho utilizamos esta técnica para a classificação automática de imagens. Como os dados de entrada para SVM's são vetores de características dos exemplos dos conjuntos de treino e teste, extraíram-se as propriedades visuais das imagens dos dois conjuntos, e com os vetores gerados desta extração foi possível utilizar a aprendizagem obtida pelo método para inferir as classes de imagens não-classificadas previamente.

3.5 Métricas de avaliação

Para comparar o método proposto com os baselines, é necessário que sejam utilizadas métricas que avaliem o desempenho dos algoritmos nas diferentes etapas. Na fase de classificação, utilizamos a métrica de Macro-Precisão. Na fase de busca utilizamos a

métrica de Precisão em N. Estas métricas são comumente utilizadas em Aprendizagem de Máquina e em Recuperação da Informação. Descreveremos a seguir cada uma destas métricas.

3.5.1 Macro-Precisão

Em um problema de classificação, uma métrica utilizada para calcular o desempenho de um algoritmo é a Macro-Precisão. Esta métrica consiste em calcular a porcentagem de acerto em cada classe c . Dado um problema com n classes, calcula-se a precisão do classificador em cada classe c_1, c_2, \dots, c_n . Em seguida, é calculada a média destas precisões, proporcionalmente ao tamanho de cada classe. Este cálculo permite que mesmo em uma base desbalanceada, todas as classes tenham a mesma importância no resultado final, independente do tamanho. Consideramos este fator, pois esperamos que o acerto seja similar tanto em classes com grande número de instâncias, quanto em classes com um pequeno número de instâncias. O cálculo da Macro-precisão pode ser visto na Figura 3.14.

$$\text{macro_precision} = \frac{1}{n_{\text{labels}}} \sum_{j=0}^{n_{\text{labels}}-1} \text{precision}_j,$$

Figura 3.14. Cálculo da macro-precisão

3.5.2 Precisão em N

Em um problema de busca, Precisão é a fração de itens recuperados que são relevantes para a necessidade de informação do usuário. A Precisão leva em conta todos os documentos recuperados. Matematicamente, a Precisão pode ser descrita conforme a Figura 3.15. Como o contexto deste trabalho é de recuperação da informação multimídia, considera-se uma imagem como um documento.

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Figura 3.15. Cálculo de precisão

Como o número de itens recuperados para uma dada consulta é grande, podemos selecionar apenas o topo do ranking com n respostas retornadas pelo sistema. Esta métrica é chamada de Precisão em n ou P@n, onde n é o número de respostas do topo do ranking.

3.5.3 MAP- *Mean Average Precision*

A medida MAP é largamente utilizada na literatura para a avaliação de sistemas de Recuperação da Informação. Esta métrica pode ser entendida como precisão média nos documentos relevantes, onde a precisão P em cada ponto é determinada quando um novo documento relevante é recuperado. O valor de P será zero para cada documento relevante não recuperado. A média das precisões encontradas nos n primeiros valores, sobre o número de documentos relevantes é o valor de MAP@n. O cálculo do MAP pode ser visto na figura 3.16. Quanto mais documentos relevantes forem recuperados próximos ao topo da resposta, maior sera o valor de MAP.

$$MAP = \frac{1}{N} \sum_{j=1}^N \frac{1}{Q_j} \sum_{i=1}^{Q_j} P(doc_i)$$

Figura 3.16. Cálculo do MAP - *Mean Average Precision*

Capítulo 4

O Método

Neste capítulo vamos apresentar nossa abordagem para melhorar os resultados da busca de imagens de produtos: método de Categorias Visuais - VisCat. Primeiramente apresentaremos uma visão geral do método, e posteriormente detalharemos cada etapa do processo.

4.1 Explorando Categorias Visuais- O método VisCat

O método VisCat pode ser dividido em dois algoritmos principais. O primeiro é a *Construção da Taxonomia Visual*, que consiste em estabelecer uma taxonomia visual - que é baseado nas características visuais das imagens. O segundo algoritmo consiste em realizar a *Busca em Escopo Reduzido* baseado na taxonomia visual, para uma dada uma imagem de consulta. Em todo o processo usamos descritores *CBIR* para extrair os vetores de características das imagens. No esquema da Figura 4.1, temos uma visão em alto nível da execução do VisCat. Na primeira etapa, a base de imagens é classificada conforme a taxonomia visual do domínio em questão. Após este passo, cada nova consulta terá uma classe atribuída, então a busca se dará apenas no conjunto de imagens da classe da consulta. Neste caso, a imagem de consulta foi prevista como uma instância da categoria N, então o algoritmo de busca será executado apenas nesta categoria.

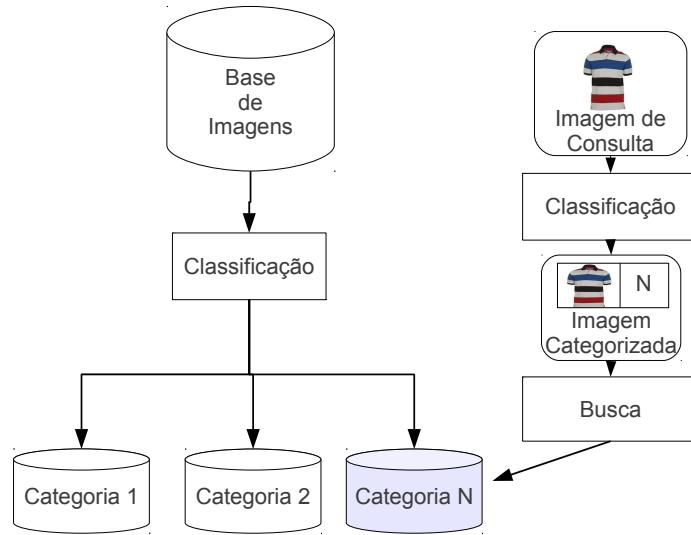


Figura 4.1. Esquema simplificado do método VisCat

4.1.1 Construção da Taxonomia Visual

Para que seja possível a construção de uma taxonomia para uma dada base de imagens, estas devem ter o mesmo padrão visual: instâncias da mesma classe visual devem ter características visuais em comum. Para que isto ocorra, a definição das próprias classes deve levar em conta a presença de informações visuais suficientes para distinção entre instâncias de classes diferentes. Na Figura 4.2, é possível visualizar um exemplo uma taxonomia visual com informações visuais que distinguem instâncias de diferentes classes.

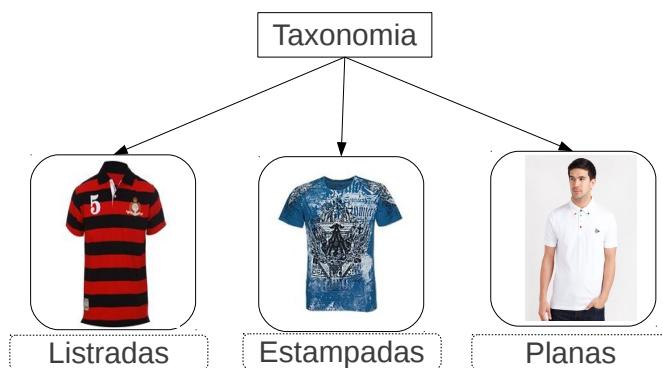


Figura 4.2. Exemplo de uma taxonomia com diferentes padrões visuais

O processo de construção da taxonomia visual possui uma etapa manual. Como as imagens da base devem ser categorizadas de forma automática, é necessário que seja fornecido um conjunto de imagens de exemplo de cada classe da taxonomia visual, para que o aprendizado seja possível. O processo completo de construção da taxonomia visual descrito no Algoritmo 1.

Algoritmo 1 Construção da Taxonomia Visual

Entrada: base de imagens I

Saída: taxonomia visual \mathcal{T}

Fase de Seeding – Construir uma taxonomia visual \mathcal{S}

- 1: **para cada** classe $s \in \mathcal{S}$ **do**
- 2: selecionar uma amostra P com ao menos n imagens para popular s
- 3: **para cada** imagem $i \in P$ **do**
- 4: *i.classe* $\leftarrow s$
- 5: $I \leftarrow I - \{i\}$
- 6: **fim para**
- 7: **fim para**

Fase de expansão – Expandir \mathcal{S} em \mathcal{T}

- 8: $\mathcal{T} \leftarrow \mathcal{S}$
- 9: **para cada** imagem $i \in I$ **do**
- 10: *i.classe* \leftarrow Classificar(i, \mathcal{T})
- 11: $I \leftarrow I - \{i\}$
- 12: **fim para**

É necessário selecionar alguns exemplos de imagens da base-alvo (Linha 2) para popular cada categoria visual (Linhas 1 a 7), assinalando manualmente uma classe para cada imagem de exemplo (Linha 4). Por exemplo, em uma base de camisas masculinas, podemos estabelecer uma taxonomia com duas classes principais (mangas longas e mangas curtas), então teríamos que escolher algumas imagens de exemplo (desta base de imagens) e classificar manualmente essas imagens em uma destas duas classes.

Uma vez que a taxonomia-semente foi construída, podemos classificar toda a base não-classificada (Linhas 8 a 12). Nesta fase, os exemplos de cada classe na taxonomia-semente serão usados na fase de aprendizagem do classificador SVM. Então, cada imagem não-classificada terá uma classe atribuída automaticamente (Linha 10)

Cada instância é descrita por um vetor de características baseado em propriedades visuais. O problema discutido neste trabalho é baseado em informação de textura de

roupas, então o EHD (MPEG7) foi o descritor escolhido para gerar os vetores de características: ao visualizar a textura como propriedade visual mais importante de uma imagem, será possível, computacionalmente, determinar que uma camisa com listras vermelhas é visualmente semelhante a uma camisa com listras azuis, pelo fato das proximidades das respectivas informações texturas. Este algoritmo foi escolhido baseado em experimentos preliminares. Como EHD é baseado fortemente em informação de textura, então imagens com o mesmo padrão visual e cores diferentes deverão ser classificados na mesma categoria visual.

Quanto ao aprendizado, o problema de classificação pode ter de ser resolvido em uma base com uma taxonomia com mais de uma classe. Dada esta particularidade, utilizamos uma implementação SVM que abstrai o problema multiclasse Crammer & Singer [2002].

4.1.2 Busca em escopo reduzido

Em nossa abordagem fazemos a busca em escopo reduzido (com categorias visuais) em detrimento da busca em geral, pois a busca em geral além de geralmente necessitar de alguma filtragem para remover possíveis ruidos da resposta, é computacionalmente mais cara: ao invés de se ter o custo de calcular similaridades de uma imagem contra uma base inteira, teremos o custo de calcular similaridades com subconjuntos de uma base. Mesmo com métodos de busca multidimensional que reduzem a quantidade de comparações, como a KD-tree (Bentley [1975]), ainda é considerável o ganho de desempenho com o corte dos candidatos em tempo de execução, onde tal ganho será diretamente proporcional ao tamanho da base.

Ao fim da classificação visual da base de imagem, é esperado que o classificador tenha aprendido como prever a classe de qualquer imagem dada como entrada, mesmo que esta imagem não esteja na coleção de treino. Esta classificação automática será usada para melhorar os resultados de uma consulta, reduzindo o escopo de busca a classe prevista para a imagem de consulta. O processo que constitui esta fase está descrita no Algoritmo 2.

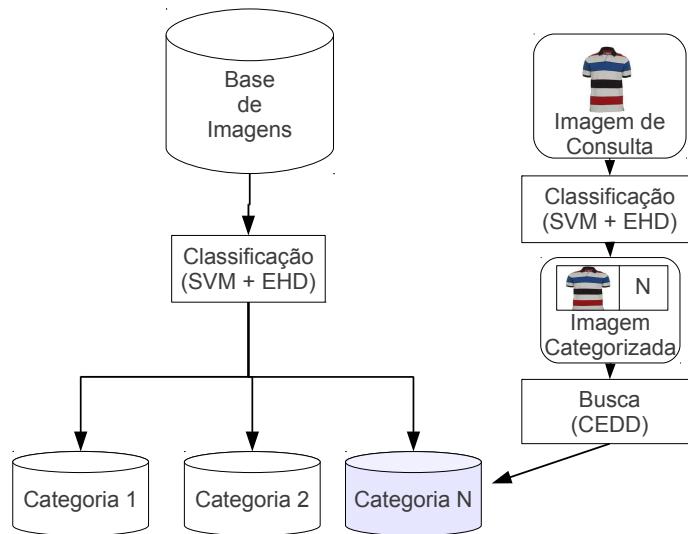
Cada nova consulta q será automaticamente classificada (Linha 1), e o algoritmo de busca irá calcular a similaridade da consulta apenas na categoria assinalada para q (Linhas 2 a 5). Para realizar a busca na categoria prevista, utilizamos o descritor CEDD, baseado em cor e textura.

Na Figura 4.3 mostramos com as etapas de classificação e busca detalhadamente, onde se aplicam os descritores de imagem EHD, CEDD e o classificador SVM.

Algoritmo 2 Busca em Escopo Reduzido

Entrada: imagem de consulta q
Entrada: taxonomia \mathcal{T} da base de imagens (c_1, c_2, \dots, c_n)
Saída: lista de imagens similares à imagem de consulta
 Redução do escopo da busca

- 1: $q.classe \leftarrow \text{Classificar}(q, \mathcal{T})$
- Execução da busca
- 2: **para cada** imagem $i \in c_j, j = q.classe$ **do**
- 3: $sim \leftarrow \text{calcular similaridade}(q, i)$
- 4: respostas $\leftarrow \text{insere ordenado}(q, sim)$
- 5: **fim para**
- 6: exibir(respostas)

**Figura 4.3.** Esquema do método VisCat

Ressalta-se que o descritor CEDD utilizado na busca, o descritor EHD utilizado na classificação, e o próprio classificador SVM podem ser substituídos por outros descritores e classificadores, conforme as peculiaridades de cada domínio de aplicação. Se o domínio da aplicação for um conjunto de imagens em tons de cinza, o descritor CEDD utilizado na busca em escopo reduzido poderia ser substituído por outro descritor, que não utilize tantas informações de cor, como o próprio EHD.

Capítulo 5

Experimentos

Os experimentos deste trabalho foram divididos em duas principais partes: classificação visual e busca visual. Na primeira parte, utilizamos uma base de 5.000 imagens de roupas masculinas. Na segunda parte, a busca visual foi executada em duas bases distintas: a primeira, com as mesmas 5.000 imagens de roupas masculinas e uma segunda, com 23.000 imagens de produtos de vestuário em geral.

5.1 Configuração do Experimento

Em nossos experimentos, utilizamos algumas ferramentas. Para a geração dos vetores de características dos descritores EHD e CEDD, utilizamos a ferramenta Img(Rummager)Chatzichristofis et al. [2009]. Esta ferramenta gera arquivos XML com estes vetores, que posteriormente foram processados por scripts para serem utilizados na fase de classificação. Para a fase de aprendizagem de máquina, utilizamos a ferramenta `svm_multiclass`¹, que abstrai o problema multiclasse para o uso de SVM. Na fase de avaliação, foi desenvolvido um sistema que permitiu que usuários selecionassem as imagens relevantes para uma dada consulta. Os experimentos foram feitos em uma máquina Pentium Dual-Core de 3.20GHz, com 2GB de memória RAM.

As imagens que compõem as bases foram coletadas entre julho e setembro de 2012, de 29 sites de comércio eletrônico distribuídos entre Brasil, Estados Unidos e Grã-Bretanha (lista completa no apêndice). Todas as consultas são imagens externas às bases, ou seja, dada uma consulta q , todas as respostas r_1, r_2, \dots, r_n são diferentes de q . Esta escolha foi feita pois uma imagem de resposta exatamente igual à consulta sempre estará no topo do ranking.

As imagens foram classificadas por 30 avaliadores, instruídos a fazer um julgamento

¹http://svmlight.joachims.org/svm_multiclass.html

binário para cada imagem, escolhendo se uma resposta era relevante ou não relevante para uma dada consulta. Cada avaliador julgou exatamente 3 respostas para cada uma das 30 consultas. Ao final da avaliação, cada resposta foi avaliada exatamente por 3 avaliadores. Sendo assim, era necessário que pelo menos 2 avaliadores julgassem uma resposta relevante para que esta fosse considerada relevante nos cálculos da métricas de avaliação. Uma tela deste sistema pode ser vista na Figura 5.1. Cada imagem da segunda linha corresponde a uma resposta para a imagem de consulta imediatamente acima.

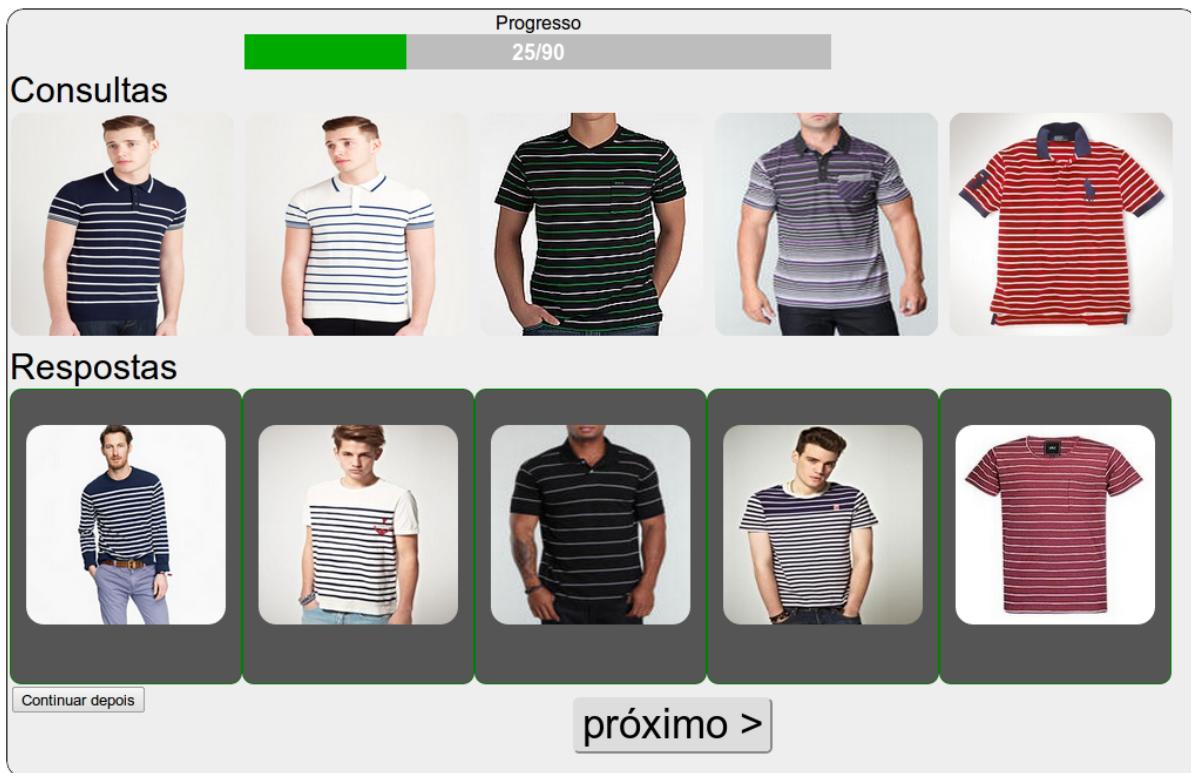


Figura 5.1. Exemplo de tela do sistema de avaliação

5.2 Classificação da base de imagens

Nesta etapa, o objetivo foi classificar automaticamente a base de 5.000 imagens. Para isso, o método deveria classificar as imagens em uma taxonomia com 5 classes relativas a vestuário masculino: Estampadas, Listradas, Planas, Planas com botão e xadrez. Estas classes foram escolhidas por serem representativas no contexto de camisas masculinas, e por sua clara distinção de conteúdo visual. A Figura 5.2 mostra a taxonomia utilizada para descrever esta base.

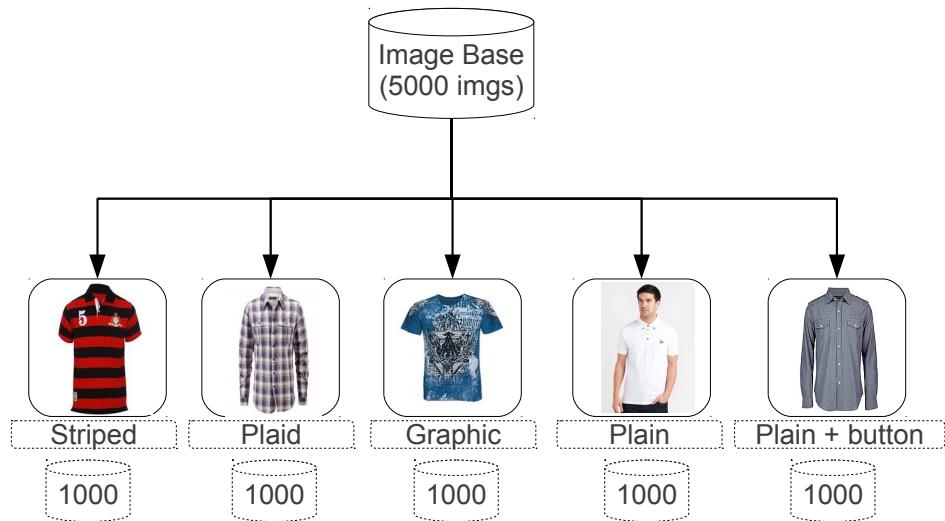


Figura 5.2. Taxonomia da base de 5.000 camisas masculinas

O número de classes se limitou a 5 pela observação da distribuição das imagens de camisas masculinas encontradas nas lojas de comércio eletrônico coletadas . Para adicionar mais classes, basta possuir mais conjuntos de imagens com outras propriedades visuais. Por exemplo: poderia ser adicionada uma categoria visual com blazers, bastaria coletar imagens deste conjunto.

Foram utilizados classificadores SVM para fazer o aprendizado e a previsão das instâncias não-classificados. A função de kernel escolhida foi a RBF (*Radial Basis Function*). Esta função foi escolhida pelo baixo número de parâmetros a ajustar.

Em experimentos preliminares, constatou-se que o descritor EHD foi eficaz na classificação das imagens nesta taxonomia. Por este motivo, avaliamos a macro-precisão obtida neste descritor, onde verificamos que a classe com maior acerto foi a xadrez. Este fato pode ser explicado pela riqueza de informações de arestas neste tipo de imagens - xadrez é um conjunto de arestas verticais e horizontais, o que se torna altamente descriptivo quando se analisa a imagem com um descritor de textura baseado em arestas, como o EHD. A Figura 5.3 contém o gráfico com a distribuição da macro-precisão entre as classes neste experimento.

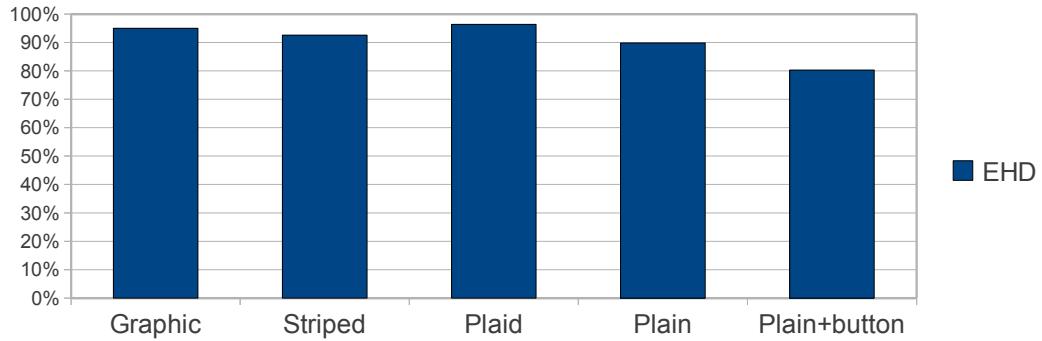


Figura 5.3. Macro-Precisão distribuída entre classes

Em seguida, realizamos experimentos para verificar o número de instâncias de treino necessárias para classificar com eficácia um conjunto não-classificado de imagens. Para esta tarefa, foram realizados experimentos com 4 descritores distintos: CEDD, EHD, BIC (Border and Interior pixel Classification)Stehling et al. [2002] e GCH(Histograma Global de Cor). Para a avaliação dos resultados, utilizamos a métrica de macro-precisão, onde o resultado (0 1), é dado pela porcentagem de acertos de previsão das instâncias em cada classe. A Figura 5.4 mostra o gráfico com a curva definida pela macro-precisão e o número de instâncias

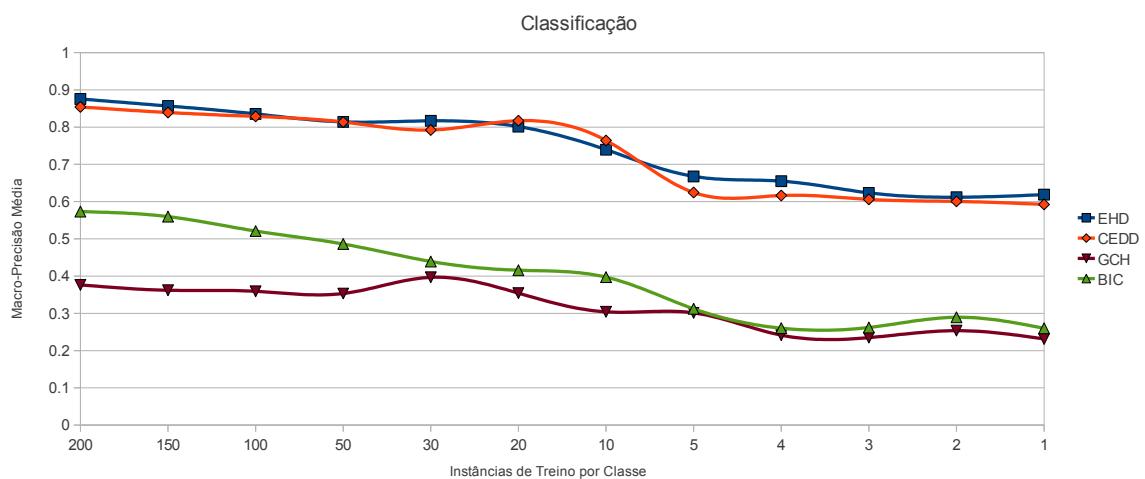


Figura 5.4. Curva macro-precisão X instâncias de teste por classe

5.3 Busca visual de imagens

Após ter as instâncias da base de imagens classificadas automaticamente na etapa anterior, foram realizadas consultas com imagens externas à base. Além disso, utilizamos uma outra base com 23.000 imagens com produtos de diversas categorias, inclusive com roupas femininas, bolsas e sapatos. O objetivo de realizar experimentos com esta segunda base é verificar se o método proposto funciona em coleções heterogêneas. Em ambas as bases foram realizadas 50 consultas, com imagens externas às bases. A métrica escolhida foi a precisão em n, com n=10, 20 e 30.

5.3.1 Coleção de Camisas Masculinas - 5.000 imagens

Neste experimento, tanto as consultas quanto as imagens da base se tratam de camisas masculinas. Este experimento foi preparado desta forma para verificar o comportamento do método em ambientes de domínio específico. A taxonomia de 5 classes foi utilizada na fase de classificação automática. As imagens desta base foram extraídas de mais de 20 diferentes sites de comércio eletrônico de roupas de vestuários espalhados pela Web. Na Figura 5.5, temos o gráfico de precisão em 10 de todas as consultas realizadas, comparando o método VisCat com o CEDD. Na Figura 5.6 temos o gráfico de precisão média geral, em 10, 20 e 30.

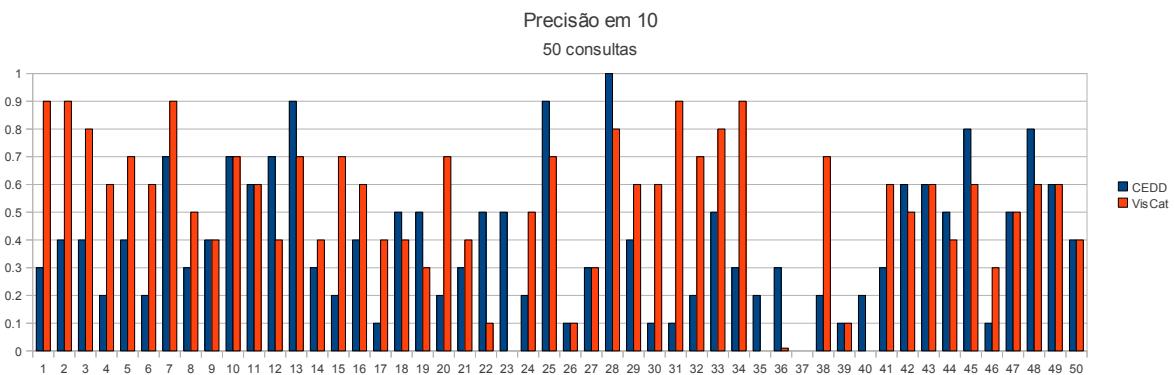


Figura 5.5. Precisão em 10 para cada consulta com a base de camisas masculinas

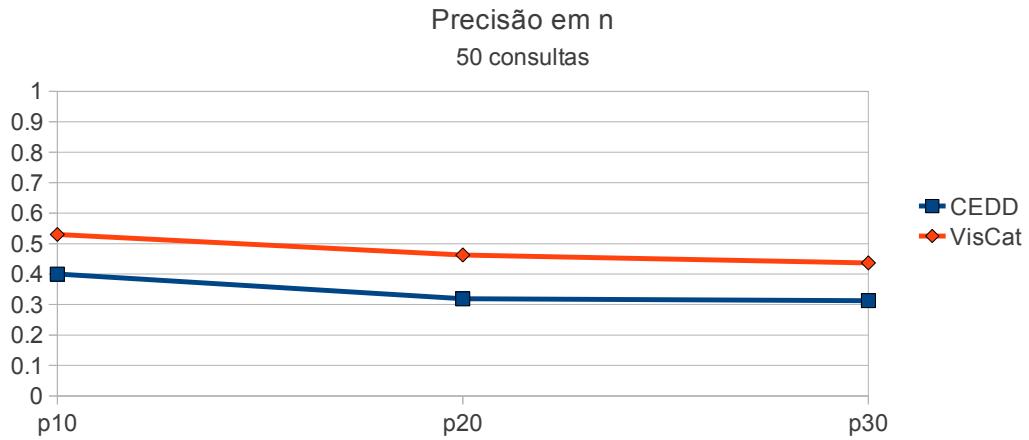


Figura 5.6. Média da precisão em 10, 20 e 30 na base de produtos de camisas masculinas

5.3.2 Coleção de Vestuário geral - 23.000 imagens

Neste experimento, tanto as consultas quanto as imagens da base se tratam de peças de vestuário em contexto geral, ou seja, roupas, calçados e acessórios, masculinos e femininos. Este experimento foi preparado desta forma para verificar o comportamento do método em ambientes de domínio amplo, sem uma taxonomia bem-definida. Porém, a mesma taxonomia de 5 classes foi utilizada na fase de classificação automática. As imagens desta base foram extraídas de mais de 2 sites de comércio eletrônico de roupas de vestuários: *Dafiti.com.br* e *Phostaus.com.br*. Na Figura 5.7, temos o gráfico de precisão em 10 de todas as consultas realizadas, comparando o método VisCat com o CEDD. Na Figura 5.8 temos o gráfico de precisão média geral, em 10, 20 e 30.

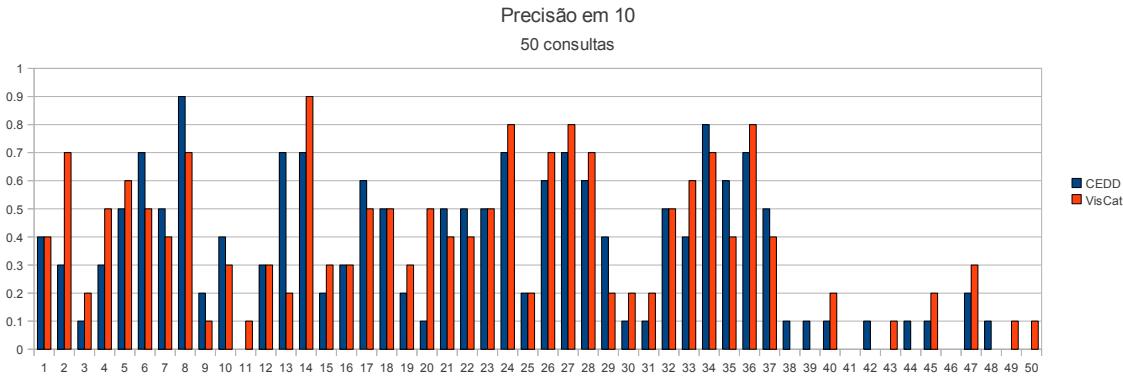


Figura 5.7. Precisão em 10 para cada consulta com a base de produtos de vestuário em geral

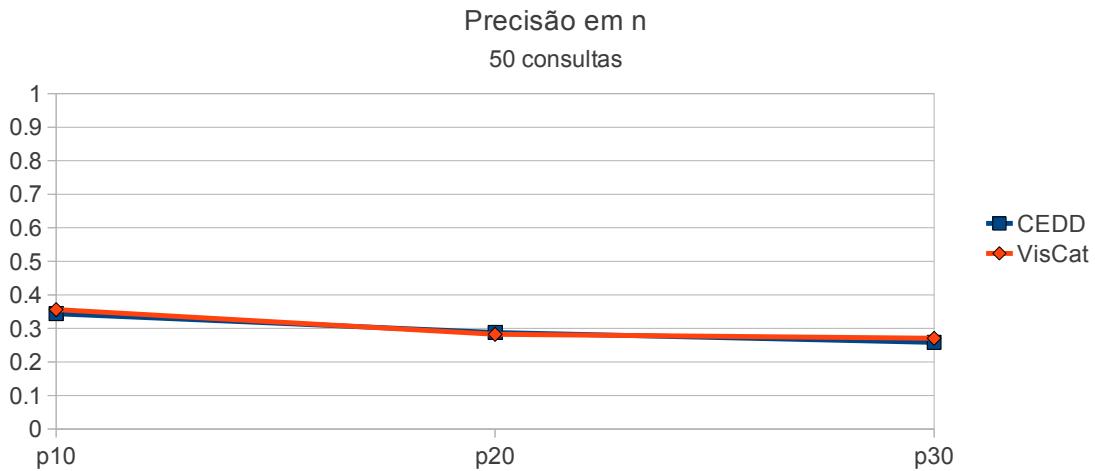


Figura 5.8. Média da precisão em 10, 20 e 30 na base de produtos de vestuário em geral

5.4 Comentários

No experimento 1 (base de 5.000 imagens) obtemos um resultado em que o uso das Categorias Visuais foi superior. Para fazer uma análise mais detalhada, devemos observar os resultados consulta-a-consulta. No gráfico mostrado na Figura 5.5, temos a seguinte distribuição: os primeiros 10 elementos são as consultas com imagens de camisas estampadas. Os próximos 10 são de camisas xadrez, planas, planas com botão e listradas, respectivamente. Verifica-se que, consulta-a-consulta, os maiores ganhos com o VisCat foram nas categorias de “estampadas” (VisCat superior em 8 consultas) e na categoria

“planas” (Viscat superior em 6 consultas). Esse resultado se explica pela ideia de que quando um usuário consulta um item de uma certa categoria visual mesmo que não tenha conhecimento da taxonomia, espera que produtos da mesma categoria visual sejam retornados , principalmente quando for uma categoria bem definida.

Para exemplificar os diferentes resultados em cada consulta, separamos dois exemplos contrapostos. No primeiro caso, quando uma consulta de uma camisa xadrez é realizada, esperam-se que camisas xadrez sejam retornadas, e camisas listradas neste caso seriam não-relevantes. Como o VisCat reduz o escopo da busca a uma categoria, há um menor número de documentos não-relevantes na resposta quando a consulta for corretamente classificada. Na segunda coluna principal da Figura 5.9 pode-se ver um exemplo prático de quando ocorre este ganho do VisCat. Adversamente a este caso, quando a consulta é classificada em uma classe visual diferente da qual deveria pertencer, apenas as imagens da classe erroneamente previstas serão retornadas. Como exemplo, temos uma consulta de uma imagem de uma camisa plana que o classificador considerou como estampada. Uma vez que isto ocorra, os resultados serão apenas imagens de camisas estampadas, as quais não foram consideradas relevantes pelos avaliadores. Este contra- exemplo pode ser visto primeira coluna principal da Figura 5.9

Exemplo 1 – CEDD maior precisão em 10 Consulta: 		Exemplo 2 – VisCat maior precisão em 10 Consulta: 	
Top 10 CEDD	Top 10 VisCat	Top 10 CEDD	Top 10 VisCat
			
			
			
			
			

Figura 5.9. Exemplos de consulta e respostas na base de Roupas Masculinas

No experimento 2 (base de 23.000 imagens) obtemos um resultado em que o uso das Categorias Visuais não trouxe resultados superiores ao da busca sem a classificação. Isto ocorreu pois a taxonomia utilizada não foi alimentada pela base em questão, mas pelas 5 classes de imagens da coleção de 5.000 itens. Sendo assim, o treinamento foi realizado com uma taxonomia que não previa outras classes além das 5 utilizadas em treinamento. O objetivo era verificar qual seria o comportamento do VisCat em um ambiente ruidoso. Com os resultados obtidos, percebe-se que o ganho do método sobre a busca visual comum é dependente da taxonomia utilizada.

Ao se fazer a análise consulta-a-consulta, destacamos também dois exemplos contrapostos. Na coluna principal da esquerda, temos um exemplo onde o VisCat obteve melhor resultado. A taxonomia utilizada para treino, apesar de não ter algumas classes como bolsas, calças e vestidos, agrupou visualmente os elementos da base de forma que apresentou resultados melhores do que os obtidos sem a classificação neste caso. Na coluna principal da direita, os dois métodos trouxeram roupas femininas como resposta, porém o resultado da busca sem classificação foi considerado superior, trazendo mais itens relevantes na resposta. Ambos exemplos podem ser vistos na Figura 5.10.

Exemplo 1 – VisCat maior precisão em 10		Exemplo 2 – CEDD maior precisão em 10	
Consulta:		Consulta:	
Top 10 CEDD	Top 10 VisCat	Top 10 CEDD	Top 10 VisCat

Figura 5.10. Exemplos de consulta e respostas na base de Vestuário em Geral

Capítulo 6

Conclusões e trabalhos futuros

O método apresentado se mostrou eficaz quanto a taxonomia visual da base era bem definida. Em um ambiente real, as informações textuais poderiam ser aproveitadas para facilitar a construção da taxonomia semente. Já na etapa de busca visual, podem ser utilizados descritores mais robustos e mais caros computacionalmente, como o SIFT (*Scale Invariant Features Transform*) Lowe [1999] e o SURF *Speeded-Up Robust Features* Bay et al. [2008], que são conhecidos na literatura por sua alta taxa de acerto Mikolajczyk & Schmid [2005]. Com a busca sendo realizada confrontando a imagem de busca com um número menor de imagens da base (busca em escopo reduzido), o tempo de processamento, que é a principal desvantagem destes descritores, seria reduzido.

Referências Bibliográficas

- Bay, H.; Ess, A.; Tuytelaars, T. & Van Gool, L. (2008). Speeded-up robust features (surf). *Comput. Vis. Underst.*, 110(3):346--359. ISSN 1077-3142.
- Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509--517. ISSN 0001-0782.
- Chatzichristofis, S. A. & Boutalis, Y. S. (2008). Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. Em *Proceedings of the 6th international conference on Computer vision systems*, ICVS'08, pp. 312--322, Berlin, Heidelberg. Springer-Verlag.
- Chatzichristofis, S. A.; Boutalis, Y. S. & Lux, M. (2009). Img(Rummager): An interactive content based image retrieval system. Em *International Workshop on Similarity Search and Applications*, pp. 151--153.
- Chen, Y.; Yu, N.; Luo, B. & Chen, X.-w. (2010). ilike: integrating visual and textual features for vertical search. Em *Proceedings of the international conference on Multimedia*, MM '10, pp. 221--230, New York, NY, USA. ACM.
- Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273--297.
- Crammer, K. & Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265--292. ISSN 1532-4435.
- Fei-Fei, L.; Fergus, R. & Perona, P. (2003). A bayesian approach to unsupervised one-shot learning of object categories. Em *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pp. 1134--, Washington, DC, USA. IEEE Computer Society.
- Fritz, M. & Schiele, B. (2008). Decomposition, discovery and detection of visual categories using topic models. Em *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1--8. ISSN 1063-6919.

- Iwayama, M. & Tokunaga, T. (1995). Cluster-based text categorization: a comparison of category search strategies. Em *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '95*, pp. 273--280, New York, NY, USA. ACM.
- Jing, Y. & Baluja, S. (2008). Pagerank for product image search. Em *Proceedings of the 17th international conference on World Wide Web, WWW '08*, pp. 307-316, New York, NY, USA. ACM.
- Kejia, W.; Honggang, Z.; Lunshao, C.; Ying, H. & Ping, Z. (2011). A comparative study of moment-based shape descriptors for product image retrieval. Em *Image Analysis and Signal Processing (IASP), 2011 International Conference on*, pp. 355-359.
- Lowe, D. G. (1999). Object recognition from local scale-invariant features. Em *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2, ICCV '99*, pp. 1150--, Washington, DC, USA. IEEE Computer Society.
- Mikolajczyk, K. & Schmid, C. (2005). A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615-1630. ISSN 0162-8828.
- Ouyang, Y. (2009). Clothes image searching system based on sift features. Em *E-Business and Information System Security, 2009. EBISS '09. International Conference on*, pp. 1-5.
- Park, D. K.; Jeon, Y. S. & Won, C. S. (2000). Efficient use of local edge histogram descriptor. Em *Proceedings of the 2000 ACM workshops on Multimedia, MULTIMEDIA '00*, pp. 51-54, New York, NY, USA. ACM.
- Stehling, R.; Nascimento, M. & Falcão, A. (2002). A compact and efficient image retrieval approach based on border/interior pixel classification. Em *Proceedings of the eleventh international conference on Information and knowledge management, CIKM '02*, pp. 102-109, New York, NY, USA. ACM.
- Sural, S.; Qian, G. & Pramanik, S. (2002). Segmentation and histogram generation using the hsv color space for image retrieval. Em *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 2, pp. II-589. IEEE.
- Tseng, C.-H.; Hung, S.-S.; Tsay, J.-J. & Tsaih, D. (2009). An efficient garment visual search based on shape context. *W. Trans. on Comp.*, 8(7):1195-1204. ISSN 1109-2750.

- Vijayanarasimhan, S. & Grauman, K. (2008). Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. Em *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8. ISSN 1063-6919.
- Xie, X.; Lu, L.; Jia, M.; Li, H.; Seide, F. & Ma, W.-Y. (2008). Mobile search with multimodal queries. *Proceedings of the IEEE*, 96(4):589–601. ISSN 0018-9219.

Anexo A

Lista de sites coletados para montagens das bases de imagens

- <http://www.dafiti.com.br/>
- <http://www.posthaus.com.br/>
- <http://www.bananarepublic.gap.com/>
- <http://www.oldnavy.gap.com/>
- <http://www.piperlime.gap.com/>
- <http://www.menswearhouse.com/>
- <http://www.forever21.com/>
- <http://www.drjays.com>
- <http://www.coggles.com/>
- <http://www.shopstyle.co.uk/>
- <http://www.menswearhouse.com/>
- <http://www.tillys.com/>
- <http://www.saksfifthavenue.com/>
- <http://www.buckle.com/>
- <http://www.bbclothing.co.uk/>

ANEXO A. LISTA DE SITES COLETADOS PARA MONTAGENS DAS BASES DE IMAGENS

- <http://www.stand-out.net/>
- <http://www.jigsaw-online.com/>
- <http://www.glamour.com.br/>
- <http://www.hackett.com/>
- <http://www.all saints.com/>
- <http://www.bensherman.com/>
- <http://www.sunspel.com/>
- <http://oliverspencer.co.uk/>
- <http://www.my-wardrobe.com/>
- <http://www.oipolloi.com/>
- <http://www.reissonline.com/>
- <http://www.folkclothing.com/>
- <http://www.howies.co.uk/>
- <http://www.cult.co.uk/>