Universidade Federal do Amazonas
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Programa de Pós-Graduação em Informática

Detecção Automática de Conteúdo Ofensivo na Web

Ruan Josemberg Silva Belém

Manaus – Amazonas
Maio de 2006

Ruan Josemberg Silva Belém

Detecção Automática de Conteúdo Ofensivo na Web

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação.

Orientador: Prof. João Marcos Bastos Cavalcanti, Ph.D.

Ruan Josemberg Silva Belém

Detecção Automática de Conteúdo Ofensivo na Web

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Departamento de Ciência da Computação da Universidade Federal do Amazonas, como requisito parcial para obtenção do Título de Mestre em Informática. Área de concentração: Recuperação de Informação.

Banca Examinadora

Prof. João Marcos Bastos Cavalcanti, Ph.D. – Orientador
Departamento de Ciência da Computação – UFAM/PPGI

Prof. Dr. Edleno Silva de Moura
Departamento de Ciência da Computação – UFAM/PPGI

Prof. Dr. Marcos André Gonçalves
Departamento de Ciência da Computação – UFMG

Manaus – Amazonas
Maio de 2006

A Família Belém.

Agradecimentos

A Deus, acima de tudo.

A minha família pelo apoio incondicional.

Ao meu orientador, João Cavalcanti e ao Prof. Laurindo Campos, pelo incentivo e por terem acreditado na minha capacidade.

Aos amigos Célia Santos, Márcio Vidal e Patrícia Peres pelos inesquecíveis momentos do quarteto.

Aos colegas Arelian Maia e Péricles Oliveira, pela grande ajuda no primeiro ano do mestrado.

Ao amigo Ícaro, pelo suporte aos experimentos e pela disponibilidade.

A Fundação de Amparo à Pesquisa do Estado do Amazonas - FAPEAM, pelo suporte financeiro.

Ao PPGI, pela oportunidade.

A todos aqueles que ajudaram de alguma forma na realização deste trabalho, o meu mais profundo agradecimento.

Impossível é tudo aquilo com que não conseguimos sonhar.

Ruan Belém

Resumo

A *World Wide Web* (*Web*) é uma fonte de informação com grande quantidade e diversidade de conteúdo, incluindo material de caráter ofensivo relacionado à pornografia. Diante deste cenário, existe a necessidade de detectar tal conteúdo ofensivo de maneira a evitar que o mesmo seja indevidamente acessado por crianças ou por funcionários de empresas, onde o acesso a este tipo de conteúdo geralmente não é permitido. Embora este tipo de informação possa estar presente na *Web* em forma de texto, vídeo ou sons, grande parte deste conteúdo está disponibilizado na forma de imagens. O problema de identificação de imagens ofensivas pode ser visto como um problema de classificação. Como as imagens em questão estão inseridas em páginas *web*, além das informações que podem ser extraídas da própria imagem, também têm-se as informações textuais encontradas nas páginas que possuem as imagens. Após a extração de evidências a classificação é realizada usando-se um classificador baseado em SVM treinado com uma coleção de 1000 imagens ofensivas e 1000 imagens não-ofensivas. Este trabalho apresenta duas abordagens diferentes para detecção de imagens ofensivas na *Web*: a primeira, baseada no conteúdo da imagem e a segunda, baseada em evidências textuais extraídas das páginas *web* onde se encontram as imagens. Ambas as abordagens se mostraram eficazes na detecção de imagens ofensivas, apesar de utilizarem algoritmos simples para a extração de informações relacionadas às imagens.

Palavras-chave: Recuperação de Informação, Detecção de imagens ofensivas, Web, Processamento Digital de Imagens.

Abstract

The World Wide Web is a huge source of diverse information, including offensive material such as pornography related content. This poses the problem of automatcally detecting offensive content as a way to avoid unauthorised access, for instance, by children or by employees during working hours. Although this sort of information is published in many forms, including text, sound and video, images are the most common form of publication of offensive content on the Web. Detecting offensive images can be considered as a classification problem. Given that images are part of Web pages, textual information can be used as important evidence along with the content extracted from images, such as colour, texture and shapes. This dissertation proposes two distinct approaches for automatic detection of offensive images on the Web. The first is based on image content, specifically colour. The second approach is based on textual terms extracted from the Web page that present the images. After evidence extraction the classification is performed using the SVM technique, based on a collection of 1000 offensive images and 1000 non-offensive images for training. Experiments carried out have shown that both approaches are effective, although they rely on simple algorithms for extracting evidences related to the images.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A *World Wide Web* (*Web*) é uma fonte importante de informação, com grande quantidade e diversidade de conteúdo. A popularização das câmeras digitais, celulares e outros dispositivos capazes de produzir imagens diretamente na forma digital, aliada a facilidade de publicação, faz com que haja grande quantidade de conteúdo multimídia disponível na Internet. Na *Web*, a liberdade de expressão é assegurada pela falta de controle do conteúdo publicado, o que contribui bastante para seu crescimento, porém isso pode representar um problema quando esta liberdade, tanto para publicação quanto para acesso, é utilizada de forma inapropriada. Existe uma grande quantidade de conteúdo ofensivo na Internet e este pode ser involuntariamente acessado por crianças ou indevidamente em instituições onde o acesso a este tipo de conteúdo não é autorizado.

Embora a Internet tenha trazido vantagens para as empresas, como o livre acesso a uma grande quantidade de informação pelos seus funcionários e o compartilhamento imediato dessas informações, existe uma preocupação em evitar que a utilização desse recurso possa diminuir a produtividade dos funcionários, causando prejuízo à empresa. Este fato já acontece em muitas empresas, onde os funcionários ocupam parte do tempo destinado às suas atividades com assuntos pessoais, tais como leitura de mensagens e acesso à páginas de conteúdos diversos, incluindo conteúdo ofensivo.

Entende-se por conteúdo ofensivo aquele relacionado à pornografia, o qual está presente na *Web* em forma de texto, imagem, vídeo ou outros formatos multimídia. Para observar conteúdo ofensivo em texto, este precisa ser lido e no caso de vídeos e animações, geralmente existe a necessidade de visualizadores específicos. Entretanto, praticamente todos os navegadores *Web*

atuais, são capazes de exibir imagens, sendo esta mídia a mais utilizada pelas facilidades de acesso e visualização. Desta forma, um classificador de imagens é um elemento importante na detecção de conteúdo ofensivo na *Web*, considerando que, em alguns casos, a presença de apenas uma imagem de nudez faz com que a página inteira seja considerada ofensiva. Outra evidência que não pode ser descartada é o texto da página onde as imagens se encontram, seja do título, apontadores ou mesmo do texto próximo às imagens, devido à formatação.

Há diversas aplicações para um sistema de detecção de imagens ofensivas, inclusive com grande valor comercial. Tal sistema pode ser aplicado como um *plug-in* em diversos sistemas comerciais, como navegadores *Web*, onde poderiam censurar automaticamente imagens consideradas inadequadas para a visualização. Outra aplicação possível seria um sistema para encontrar pornografia em estações de trabalho de instituições, tais como empresas e universidades. Nestas instituições, um filtro poderia ser acoplado ao servidor *proxy*[1] para evitar o acesso de conteúdo pornográfico pelos usuários. As máquinas de busca na *Web* podem utilizar este tipo de classificador para filtrar páginas indesejadas nos resultados apresentados aos seus usuários.

1.1 Detecção de Imagens Ofensivas

O problema de identificação de imagens ofensivas pode ser visto como um problema de classificação. Como as imagens em questão estão inseridas em páginas *Web*, além das informações que podem ser extraídas da própria imagem, também têm-se as informações textuais encontradas nas páginas que possuem as imagens. Neste trabalho são apresentadas duas abordagens para extração de características e classificação das imagens: a primeira, baseada em evidências de conteúdo da imagem, especificamente cor. A segunda baseada nas evidências textuais. Também é verificado se a precisão do classificador pode ser otimizada pela combinação dos resultados individuais de cada abordagem.

Abordagens recentes de detecção de imagens ofensivas [17, 20, 18] utilizam uma combinação de evidências baseadas em características como cor, textura, forma, entre outras de baixo nível, para descrever o conteúdo das imagens e realizar comparações que permitam sua classificação de forma precisa. Outras abordagens são baseadas em texto ou ainda em um sistema de rótulos

[1] Um servidor que atua entre uma aplicação cliente, como um navegador *Web*, e um servidor real, interceptando as requisições direcionadas à este e verificando se pode atendê-las ele próprio ou, em caso negativo, as repassa para o servidor real.

de conteúdo [12]. Entretanto, o uso de muitas características na descrição das imagens pode representar um aumento indesejado na complexidade e no tempo de processamento dos algoritmos de classificação. Para evitar este problema, neste trabalho propõem-se algoritmos simples, mas eficientes na representação das características das imagens, que trabalham, por exemplo, apenas com a característica de cor. Existe uma associação clara entre as imagens ofensivas e a quantidade de *pixels* de cor de pele presentes nas mesmas, pois a situação mais comum neste tipo de imagem é a presença de nudez e portanto grandes áreas de pele expostas. Além das informações de conteúdo da imagem, o texto presente nas páginas *Web* onde as imagens ofensivas se encontram é bem característico e também pode ser usado como fonte de evidência na detecção de imagens ofensivas. Como exemplos, têm-se: o título da página, o texto alternativo associado a cada imagem quando esta não pode ser exibida, passagens de texto nas proximidades da localização da imagem na página e o conteúdo dos apontadores (*links*).

De posse de um conjunto de características que descrevem a imagem, existe a necessidade de prover uma forma de usar estas informações para detectar a presença de conteúdo ofensivo nas mesmas. Isso pode ser feito através de um classificador, utilizando, por exemplo, técnicas de aprendizagem de máquina, tais como árvores de decisão, máquinas de vetores suporte (*Support Vector Machines* - SVM) ou redes neurais. Estas técnicas permitem que, através de um processo de treinamento, o classificador seja capaz de identificar de forma automática e com precisão, se determinada imagem é ofensiva ou não-ofensiva.

1.2 Objetivos do Trabalho

O objetivo deste trabalho é propor duas abordagens diferentes para identificação de conteúdo ofensivo na *Web*: a primeira, baseada no conteúdo intrínseco da imagem, ou seja, as cores presentes na mesma e, a segunda abordagem, baseada em evidências textuais extraídas das páginas *Web* onde se encontram as imagens. Nas duas abordagens propostas, foram utilizados algoritmos simples para extração de informações relacionadas às imagens, apresentando resultados compatíveis com os obtidos por soluções existentes na literatura e com melhor desempenho devido a sua simplicidade. Ainda foi verificada também a possibilidade de combinar os resultados individuais de cada abordagem em um único classificador com o objetivo de melhorar a precisão.

1.3 Contribuições

Uma das principais contribuições deste trabalho é a proposta de uma abordagem para detecção de imagens ofensivas baseada em conteúdo, considerada mais simples por utilizar apenas uma característica (cor), e que ainda assim apresenta resultados satisfatórios quando comparada com abordagens existentes na literatura. Outra contribuição importante é a constatação de que, apesar de alguns autores [9, 1] afirmarem que texto não é uma fonte de evidência eficaz na detecção de conteúdo ofensivo, dependendo da forma como esta evidência é utilizada, pode apresentar os melhores resultados. De fato, no caso da Web, onde quase sempre existe texto associado às imagens, o uso de evidências de texto para detectar imagens ofensivas se mostrou bastante eficiente conforme apresentado na Seção 4.3.

1.4 Organização da dissertação

Esta dissertação está organizada em cinco capítulos, dos quais este é o primeiro. No Capítulo 2 são apresentados os principais conceitos necessários ao entendimento deste trabalho e trabalhos correlatos encontrados na literatura . Os capítulos 3 e 4 descrevem respectivamente as abordagens baseada em conteúdo e baseada em texto, apresentando como é realizada a extração de características de conteúdo da imagem ou textuais, e como estas são utilizadas para classificar imagens ofensivas. Nestes mesmos capítulos, são apresentadas informações sobre os experimentos realizados, tais como dados sobre as bases de imagens e páginas *Web* utilizadas e configurações do ambiente de experimentação e ferramentas de *software* utilizadas. No Capítulo 5, são discutidos e analisados resultados obtidos nos experimentos apresentados nos capítulos 3 e 4. Finalmente, o Capítulo 6 exibe um resumo dos resultados e conclusões obtidas e apresenta sugestões de trabalhos futuros.

Capítulo 2

Conceitos Básicos e Trabalhos Correlatos

Neste capítulo são apresentados alguns conceitos necessários ao entendimento deste trabalho. Além disso, são apresentados alguns trabalhos relacionados, que são utilizados como base de comparação com as abordagens aqui propostas, em relação a eficácia na detecção de imagens ofensivas.

2.1 Conceitos Básicos

Nesta seção são apresentados conceitos da área de processamento digital de imagens, aprendizagem de máquina e métricas de avaliação, que são utilizados nesta dissertação.

2.1.1 Imagens Digitais

Alguns conceitos importantes sobre imagens digitais, que serão utilizados no decorrer deste trabalho, são apresentados a seguir[10]:

- *Imagem Digital*: uma imagem é uma função bi-dimensional $f(x, y)$, onde x e y são coordenadas espaciais (planas). O valor de f nos pares de coordenadas é chamada de *intensidade* naquele ponto. A imagem é digital quando os valores de x, y e os valores de f são todos finitos e de quantidades discretas.

- *Pixel*: uma imagem digital é composta por um número finito de elementos, cada um com uma localização particular e um valor. Estes elementos são conhecidos como *pixels*.

- *Espaço de cor RGB*: imagens coloridas são armazenadas em três componentes primários, formando um espaço de cor. O espaço RGB é composto das cores primárias vermelho (R), verde (G) e azul (B). Estas cores são chamadas de primárias aditivas, pois são adicionadas para produzir as outras cores, como mostra a Figura 2.1. A representação pode variar, sendo a mais comum a representação de 24 *bits*, 8 para cada cor. Uma imagem digital RGB com 24 *bits* para representar a informação de cor, pode ter 2^{24} (16.777.216) combinações de (r, g, b) onde $0 \leq r \leq 255$, $0 \leq g \leq 255$ e $0 \leq b \leq 255$.

Figura 2.1: Espaço de cor RGB: Cores produzidas pela adição de cores primárias.

- *Quantização de Cor*: é a redução do espaço de cores de uma imagem, através da escolha de subconjuntos das cores originais para aproximação com as cores do novo espaço. Existe portanto um mapeamento das cores originais para as cores aproximadas no novo espaço. A quantização é uniforme, quando o espaço de cores original é dividido em subespaços com o mesmo tamanho e, não-uniforme, quando os subespaços possuem tamanho variável e a decisão sobre estes subespaços depende das cores originais. Como exemplo, na Figura 2.2 a mesma imagem é exibida em três diferentes quantizações. A quantização tem impacto direto sobre os requisitos de espaço de armazenamento das informações de cor extraídas da imagem.

- *Histograma de Cor*: distribuição estatística das cores em uma imagem. Formalmente pode ser definido como $h_{A,B,C}(a, b, c) = N \cdot Prob(A = a, B = b, C = c)$, onde A, B e C representam os três canais de cores RGB e N é o número de *pixels* na imagem. Com-

Figura 2.2: Quantização: (a) imagem original com 16.7 milhões de cores, (b) imagem quantizada em 8 cores e (c) imagem quantizada em 2 cores.

putacionalmente, consiste na discretização do número de cores através da quantização e na contagem do número de *pixels* de cada cor.

Figura 2.3: Histograma de Cor: (a) imagem digital com 9 *pixels* (b) histograma.

- *Segmentação*: é o particionamento de uma imagem nos seus diversos componentes constituintes, realizando a extração dos objetos de interesse na imagem para processamento posterior. Na prática, a segmentação consiste na classificação de cada *pixel* como sendo de uma das partes da imagem, como por exemplo, pertencente ao fundo ou ao objeto principal. Na Figura 2.4, a imagem de um avião é segmentada e pelo menos três objetos são identificados: o avião, a sombra do avião e o fundo.

- *Detecção de Borda*: uma borda é o contorno entre objetos ou entre um objeto e o fundo, indicando o limite entre objetos sobrepostos. Computacionalmente, as bordas são regiões da imagem onde ocorre uma mudança de intensidade em um certo intervalo do espaço. Os algoritmos de segmentação para detecção de borda são baseados no processo de localização

e realce dos *pixels* de borda, aumentado o contraste entre a borda e o fundo. Este processo verifica a variação dos valores de luminosidade de uma imagem. Na Figura 2.4, parte (c), pode ser observado o resultado da aplicação de um algoritmo de detecção de borda.

Figura 2.4: Segmentação: (a) imagem original (b) imagem segmentada (c) borda detectada

2.1.2 Support Vector Machines

Support Vector Machines (SVM) [8] é uma técnica de aprendizagem de máquina supervisionada para classificação, onde são necessários exemplos previamente identificados para construir um modelo. Neste método, as características dos objetos a serem classificados são transformadas em vetores de valores reais, onde cada dimensão deste vetor corresponde à uma característica. Estes vetores de características são então mapeados em um espaço com alta dimensionalidade através de um mapeamento não linear, para que neste novo espaço seja encontrado, de forma mais fácil que no espaço original, um hiperplano ótimo que separe os vetores mapeados em suas diferentes classes.

Uma função linear de decisão é definida pelo hiperplano ótimo construído neste espaço de características, de maneira que, dado um vetor de classe desconhecida, sua classe seja decidida pela aplicação desta função. A margem entre os vetores das classes é maximizada para permitir uma generalização, isto é, gerar um modelo o mais genérico possível de maneira a reduzir a taxa de erro na classificação de novos objetos de classe desconhecida. Os vetores que definem a máxima margem são chamados vetores de suporte (*support vectors*). Esses elementos são exemplificados na Figura 2.5, onde quadrados e círculos representam vetores de características dos membros de duas classes, em um problema linearmente separável. Apenas duas dimensões foram utilizadas para simplificar o exemplo. Finalmente, a classe de vetores de características não identificados pode ser predita aplicando-se o modelo gerado.

Dois parâmetros devem ser considerados quando tratamos com classificadores SVM, a função *kernel* e o parâmetro de regularização [13]. A função *kernel* calcula a similaridade entre dois

Figura 2.5: SVM separando duas classes em um espaço bi-dimensional. Os quadrados e círculos preenchidos são os vetores de suporte. (a) Hiperplano ótimo e (b) Margem ótima.

vetores de suporte e também é usada para mapeá-los no espaço com alta dimensionalidade citado anteriormente. Isto permite separar os vetores de características das diferentes classes, o que freqüentemente não é possível no espaço de entrada original. Os tipos básicos de *kernel* são: linear, polinomial, sigmoidal e funções de base radial (RBF) [5]. O parâmetro de regularização é usado para ajustar a rigidez do procedimento de otimização.

2.1.3 Métricas de avaliação

A seguir são apresentadas algumas métricas utilizadas na avaliação dos sistemas de detecção de imagens ofensivas: precisão, revocação e medida F.

Precisão e Revocação

Precisão e Revocação são métricas de qualidade usualmente utilizadas para avaliar sistemas de Recuperação de Informação. Em [15] estas medidas também são utilizadas para avaliar resultados de classificação automática de páginas *Web*. Considerando R como o conjunto de documentos relevantes identificados pelos especialistas e S o conjunto de documentos retornados pelo sistema avaliado, podem ser definidas as seguintes medidas:

Precisão é a porcentagem de documentos em S, retornados pelo sistema, que são relevantes, como definido na Equação (2.1):

$$\text{Precisão} = \frac{|R \cap S|}{|S|} \tag{2.1}$$

Revocação é a porcentagem de documentos relevantes, que estão em R, e foram retornados pelo sistema, como definido na Equação (2.2):

$$\text{Revocação} = \frac{|R \cap S|}{|R|} \tag{2.2}$$

Neste trabalho, essas duas medidas são utilizadas para avaliar a qualidade dos resultados gerados pelos sistemas de detecção de imagens ofensivas. Estes sistemas são, de fato, um tipo de classificador, onde as imagens podem ser classificadas apenas como "ofensiva"ou "não-ofensiva".

Neste caso, para avaliar tais sistemas, deve-se definir dois conjuntos de imagens relevantes, um conjunto de imagens ofensivas ($R_{ofensivo}$) e um conjunto de imagens não-ofensivas ($R_{nao-ofensivo}$), que são criados manualmente. Estes dois conjuntos são combinados, formando um conjunto A, de tamanho $|A| = |R_{ofensivo}| + |R_{nao-ofensivo}|$, que é submetido ao sistema de detecção de imagens ofensivas.

O sistema avaliado define cada uma das imagens em A como ofensiva ou não-ofensiva, isto é, o sistema divide as imagens de A em dois conjuntos: $S_{ofensivo}$ e $S_{nao-ofensivo}$. Sendo assim, após a execução do sistema de detecção de imagens ofensivas, têm-se quatro conjuntos de imagens:

- Conjunto de imagens ofensivas relevantes: $R_{ofensivo}$;
- Conjunto de imagens ofensivas retornadas pelo sistema: $S_{ofensivo}$;
- Conjunto de imagens não-ofensivas relevantes: $R_{nao-ofensivo}$;
- Conjunto de imagens não-ofensivas retornadas pelo sistema: $S_{nao-ofensivo}$.

Para cada classe de imagens, ofensiva ou não-ofensiva, computa-se as medidas de precisão e revocação, conforme as equações 2.1 e 2.2. É importante citar que em trabalhos específicos sobre detecção de imagens ofenisvas encontrados na literatura [17], é comum a utilização dos termos sensibilidade e especificidade. *Sensibilidade* é a revocação das imagens ofensivas e *especificidade* é a revocação das imagens não-ofensivas e estão definidas nas equações 2.3 e 2.4. A partir deste ponto os termos sensibilidade e especificidade serão utilizados na apresentação de resultados obtidos pelos sistemas de detecção de imagens ofensivas.

$$\text{Sensibilidade} = \frac{|R_{ofensivo} \cap S_{ofensivo}|}{|R_{ofensivo}|} \tag{2.3}$$

$$\text{Especificidade} = \frac{|R_{nao-ofensivo} \cap S_{nao-ofensivo}|}{|R_{nao-ofensivo}|} \tag{2.4}$$

É importante ressaltar que a revocação de uma classe está diretamente relacionada a precisão da oura classe, logo, as medidas de sensibilidade e especificidade podem ser definidas também em função da precisão.

Medida *F*

A *medida F* ou *média harmônica* é utilizada quando deseja-se combinar os valores de Precisão (P) e Revocação (R) em um único valor que mensure a qualidade dos resultados de um sistema [2]. A medida F é definida pela Equação 2.5:

$$F = 2 \times \frac{P \times R}{P + R} \tag{2.5}$$

Esta métrica assume valores entre 0 e 1, sendo 0 no caso em que nenhum documento foi retornado e 1 quando o sistema retorna todos os documentos relevantes com precisão máxima.

No caso dos sistemas de detecção de imagens ofensivas, é computada a medida F para cada uma das duas classes de imagens resultantes: imagens ofensivas ($F_{ofensivo}$) e imagens não-ofensivas ($F_{nao-ofensivo}$). Em seguida, calcula-se a média entre estas duas medidas, gerando o que chamou-se aqui de *medida* F *geral* (Equação 2.6):

$$\text{Medida } F \text{ Geral} = \frac{F_{ofensivo} + F_{nao-ofensivo}}{2} \tag{2.6}$$

A medida F geral representa a acurácia do classificador, ou seja, mede a qualidade dos resultados da classificação como um todo, considerando as duas classes.

2.2 Trabalhos Correlatos

Nesta seção são apresentados alguns métodos propostos para identificar imagens ofensivas. Para os métodos baseados nas características de conteúdo da imagem, os resultados dos experimentos serão apresentados utilizando os conceitos de sensibilidade e especificidade anteriormente definidos. Embora sejam apresentados dados de desempenho dos referidos métodos, estes dados devem ser avaliados levando-se em consideração os recursos computacionais disponíveis na época

da publicação dos artigos. Como a diferença chega a alguns anos em determinados casos, esses dados de desempenho devem ser considerados apenas como informativos e não para efeito de comparação.

A abordagem proposta por Forsyth and Fleck (1996) [9] é baseada na combinação das propriedades de cor e textura para filtrar regiões de pele das imagens. O algoritmo usa restrições geométricas na estrutura humana para tentar agrupar as regiões de pele em figuras humanas, ou seja, verifica o posicionamento relativo entre elementos que possivelmente representam partes do corpo humano como braços em relação ao tronco. Como relatado em [9, 17] os resultados obtidos foram 52,2% de sensibilidade e 96,6% de especificidade para um conjunto de 138 imagens de teste ofensivas, com pessoas nuas e 1401 imagens não-ofensivas. O algoritmo de agrupamento especializado leva aproximadamente 6 minutos para processar uma imagem que tenha passado pelo filtro de pele em uma estação de trabalho típica.

Wang et al (1998) [17] propôs o *Wavelet Image Pornography Elimination* (WIPE $_{TM}$) o qual utiliza uma combinação de filtros em uma seqüência, com os filtros mais rápidos sendo executados primeiro, para que as imagens não-ofensivas passem mais rapidamente reduzindo o tempo de classificação. Depois que a imagem que está sendo verificada passa por todos os filtros iniciais, o algoritmo produz vetores de características usando análise de forma, textura, histograma de cor e estatística para verificar a similaridade da mesma com um conjunto de imagens de treinamento pré-classificadas, com 500 imagens ofensivas e 8000 imagens não-ofensivas. Se a imagem de consulta "casa"com algumas imagens ofensivas entre as 15 mais similares, ela é classificada como ofensiva. O sistema processa uma consulta em menos de 10 segundos em uma estação de trabalho SUN Sparc-20, com 96% de sensibilidade e 91% de especificidade em experimentos usando um conjunto de teste com 1076 imagens ofensivas e 10809 imagens não-ofensivas.

Arentz e Olstad (2004) [1] apresentam um método baseado no conteúdo das imagens para classificar *sites Web* adultos. O algoritmo para classificação de imagens inicia com um filtro que remove os pixels que não pertencem a regiões de pele na imagem, rejeitando pixels fora de uma faixa de cor previamente determinada. Os pixels remanescentes são agrupados em regiões conectadas e identificadas chamadas de objetos. As identificações e o número de pixels de cada objeto são armazenados em vetores distintos. Estes vetores são ordenados e os N maiores objetos são analisados para extrair vetores de características compostos de cor, textura, forma, tamanho e posicionamento de cada objeto. A importância dos diferentes elementos no vetor de

características é calculada por um algoritmo genético. Os objetos analisados são classificados e a probabilidade de uma imagem ser ofensiva é calculada com base na probabilidade de cada objeto da imagem analisado pertencer a uma determinada classe de objetos. O conjunto de teste foi composto de 500 imagens ofensivas e 800 não-ofensivas. Nos experimentos relatados, em média, 11 imagens são processadas em 1 segundo em uma estação de trabalho com 1 processador, apresentando como resultados, 89,4% de precisão geral, 95% sensibilidade e 88% de especificidade.

Zhu et al (2004) [20] propõe um método adaptativo para detecção de pele e analisam o impacto das melhorias obtidas na aplicação do método para filtragem de imagens ofensivas. A detecção é realizada em duas etapas. Na primeira, pixels similares à pixels de pele são identificados utilizando um modelo de pele genérico. Na segunda etapa, o algoritmo padrão de Expectation-Maximization (EM) [4] é utilizado para treinar um Gaussian Mixture Model (GMM) [4]. Em seguida, um classificador SVM utiliza características de espaço, forma e parâmetros Gaussianos para identificar os componentes Gaussianos de pele corretos a partir do GMM treinado. Em comparação com os métodos existentes, houve uma redução no número de falsos positivos. O método foi utilizado para substituir o modelo de pele genérico do sistema de filtragem de conteúdo ofensivo MORF [19]. O componente principal do MORF é um classificador SVM para identificar imagens ofensivas onde vetor de características com 144 posições é utilizado. Estas características são baseadas em cor como histograma e cor média e textura, em três orientações diferentes (vertical, horizontal e diagonal). Foram utilizadas 13500 imagens ofensivas e a mesma quantidade de imagens não-ofensivas para treinamento e 1500 ofensivas e a mesma quantidade de não-ofensivas para teste. Utilizando um classificador hierárquico em 2 níveis, o melhor resultado obtido foi de 94.63% de precisão geral.

A principal diferença entre estes trabalhos correlatos e as abordagens propostas nesta dissertação é que estes utilizam um número maior de características e, na maioria dos casos, com algoritmos de maior complexidade para a extração destas características. O trabalho que apresenta maior semelhança é o de Zhu et al, mais especificamente o classificador MORF alterado, o qual também utiliza um classificador SVM. Porém, mesmo neste caso, um número maior de características é utilizado, o que significa um vetor com mais posições. As abordagens aqui apresentadas procuram utilizar algoritmos simples para a extração de características. No caso da abordagem baseada em conteúdo, um número reduzido de características e conseqüentemente

de posições no vetor são utilizados pelo classificador SVM.

Coelho (2004) [7] apresenta uma abordagem para busca de imagens onde a ordenação dos resultados(*ranking*) é realizada utilizando-se múltiplas evidências textuais. Foram avaliados quais partes de uma página *Web* podem ser utilizadas para compor uma descrição efetiva das imagens. Também é proposto um modelo de busca de imagens, o qual utiliza redes bayeseanas para combinar as ordenações baseadas nas múltiplas evidências em uma única ordenação de imagens. Nos experimentos foi avaliado o impacto da utilização de quatro fontes distintas de evidências associadas com as imagens na página *Web*: *tags* descritivas (`Alt`, `<IMG>` e *tags* de apontadores), *tags* de meta informações (`<TITLE>`, `<META>`), o texto completo da página e passagens de texto nas proximidades das imagens. A combinação das *tags* descritivas com passagens de 40 termos é a que apresenta os melhores resultados. Os experimentos descritos mostram um alto ganho na qualidade dos resultados das buscas de imagens quando as diferentes fontes de evidências textuais são combinadas de forma apropriada através de uma função de ordenação. Apesar deste trabalho não tratar de imagens ofensivas, o conjunto de evidências textuais que apresentou os melhores resultados, serviu como base para a abordagem baseada em texto apresentada nesta dissertação.

Capítulo 3

Detectando Imagens Ofensivas Usando Conteúdo de Imagem

A combinação de um conjunto de características tais como forma, cor e textura, tem sido a base da maioria das abordagens mais recentes para identificação de imagens ofensivas. De acordo com Arentz e Olstad [1], a chave para identificar nudez é encontrar regiões com cor de pele nas imagens. Apesar destes fatos, nesta primeira abordagem busca-se uma técnica mais simples, que utilize apenas a característica de cor, visando melhor desempenho, combinada com um classificador SVM.

Nas duas abordagens propostas neste trabalho, a técnica de aprendizagem de máquina escolhida para implementar o classificador foi o SVM. Sendo assim, podemos descrever o processo geral de construção do classificador em três etapas:

1. *Extração de Características*: Durante esta fase, um conjunto pré-definido de características representativas do conteúdo da imagem deve ser extraído, seja da própria imagem ou do texto à ela relacionado. Essas características serão transformadas em vetores de valores reais, um vetor para cada imagem. Inicialmente duas bases distintas são processadas, treinamento e teste, as quais serão utilizadas respectivamente para criação e validação do modelo de classificação.

2. *Treinamento*: A base de treinamento contém exemplos positivos e negativos que são utilizados pelo SVM para criação do modelo de classificação. Nesta fase, o SVM utiliza os

exemplos fornecidos para encontrar uma função que seja capaz de classificar novos exemplos apresentados ao classificador, dos quais se desconhece a classe.

3. *Teste e Classificação*: Após a criação do classificador, este precisa ser testado para avaliar sua eficácia, o que é feito através da comparação da classificação da base de teste feita pelo classificador com a classificação manual previamente realizada.

3.1 SNIF - Simple Nude Image Finder

A abordagem aqui proposta, chamada de *Simple Nude Image Finder* [3], ou simplesmente SNIF, tem como principal característica a simplicidade em se utilizar apenas uma fonte de evidência para representar o conteúdo das imagens. A suposição é que quanto mais simples for o método, melhor será seu desempenho em relação ao tempo de execução. A seguir, são apresentados detalhes sobre os componentes do SNIF e o seu funcionamento.

3.1.1 Extração de Características

Para extração da característica de cor das imagens utilizou-se o BIC (*Border / Interior Pixel Classification*) [16]. O BIC é composto por um algoritmo de detecção de pixels de borda/interior e uma representação compacta das características visuais extraídas das imagens (histograma de cor), obtida através de uma função logarítmica de normalização, que reduz a necessidade de espaço de armazenamento.

A análise das imagens é realizada no espaço de cor RGB, quantizado originalmente em 64 cores. A quantização pode variar sendo definida em Q cores (32,64,128). Após a quantização, cada *pixel* é classificado como (1) *borda*, quando localizado efetivamente na borda da imagem ou quando um dos seus quatro *pixels* vizinhos (cima, baixo, direita, esquerda) é de uma cor quantizada diferente ou (2) *interior*, se é da mesma cor que seus quatro vizinhos. Baseado nessa classificação, dois histogramas gerais de cor são computados. Um considerando apenas os *pixels* de borda e outro apenas os *pixels* de interior, cada um com Q posições. Uma posição contém inicialmente a quantidade total de *pixels* da respectiva cor, normalizada para valores inteiros no intervalo $[0, 255]$. Para evitar o problema no qual valores altos em uma das posições seja dominante no cálculo da distância entre dois histogramas, os valores são normalizados para uma escala logaritmica segundo a função $f(x)$, a seguir:

$$f(x) = \begin{cases} 0, & \text{se x} = 0 \\ 1, & \text{se } 0 < \text{x} \leq 1 \\ \lceil \log_2 x \rceil + 1, & \text{caso contrário} \end{cases} \tag{3.1}$$

A função $f(x)$, definida na Equação (3.1), utiliza uma escala logarítmica que reduz em 28 vezes a distância entre o menor e o maior valor possível em cada posição do histograma, dado que o valor de cada posição, originalmente entre 0 e 255, é normalizado para valores entre 0 e 9 ao ser aplicada esta função.

Como resultado deste processo, juntando os dois histogramas, têm-se um histograma com $2 \times Q$ posições, onde cada posição contém valores inteiros entre 0 e 9. Usando esse histograma, cada imagem pode ser representada por um vetor com o mesmo número de posições, $2 \times Q$. A Figura 3.1 mostra um exemplo de uma imagem e suas correspondentes imagens com *pixels* de borda e de interior obtidas com o BIC.

Figura 3.1: BIC: Imagem Original, imagem de borda e imagem de interior.

Note que, no caso de imagens ofensivas, uma grande quantidade de *pixels* de interior são de áreas de pele humana. Isso ocorre na grande maioria das imagens que contém nudez e é uma informação determinante para detecção de imagens ofensivas.

Após a extração das características de conteúdo feita pelo BIC, o formato de saída do mesmo, um histograma com $2 \times Q$ posições, é facilmente adaptado para servir como entrada do classificador SVM, o que é feito apenas acrescentando o índice de cada posição e suprimindo as posições que contém valores iguais a 0. Outra importante característica é a economia de espaço em memória obtida pelo método. Como cada posição, contém um valor inteiro no intervalo $[0, 9]$, dois valores podem ser armazenados em um byte. Assim, a informação de conteúdo de

uma imagem ocupa apenas Q bytes de espaço de armazenamento. Experimentos descritos na Seção 3.2 mostram que o histograma dos *pixels* de borda pode ser descartado sem que haja uma perda considerável na eficácia do classificador. Como conseqüência, o histograma antes com $2 \times Q$ posições, passa a ter Q posições, referentes aos *pixels* de interior. Isso reduz o espaço de armazenamento anteriormente requerido, de Q bytes para $Q/2$ bytes, ou seja, a metade do espaço de armazenamento originalmente necessário.

Tomando como exemplo a configuração de 128 cores quantizadas, utilizando-se os histogramas de borda e de interior, são usadas 256 posições armazenadas em 128 bytes. Usando apenas o histograma de interior com 128 posições, apenas 64 bytes são utilizados para armazenar o vetor de características extraído de cada imagem.

3.1.2 Treinamento usando Cor

Para a realização do treinamento é necessário que exista uma base de imagens de exemplo, tanto positivos como negativos quanto à presença de conteúdo ofensivo. Depois de processada pelo BIC, a base é convertida em vetores de características, os quais são utilizados como entrada para o processo de treinamento do classificador SVM. Como a informação contida nos vetores é relativa à quantidade de *pixels* de cada uma das diferentes cores quantizadas presentes na imagem, a suposição é que o classificador é capaz de "aprender"que possíveis imagens ofensivas possuem uma grande quantidade de *pixels* em determinadas posições do vetor. Muito provavelmente essas posições estão relacionadas com as tonalidades de cor da pele humana. Como resultado do processo, um modelo de classificação é obtido, o qual será utilizado para predizer a classe de novas imagens apresentadas ao classificador. Esse modelo será validado na fase de teste.

3.1.3 Teste e Classificação

Para validar o modelo criado no treinamento do SVM, um conjunto de imagens de teste é classificado para avaliar a eficácia do classificador. Esse conjunto é independente do conjunto de treinamento, ou seja, não existem imagens comuns aos dois conjuntos. O conjunto de teste, da mesma forma que o de treinamento, também é pré-classificado manualmente para que seja feita a comparação entre a classificação humana e a fornecida pelo classificador SVM. Após a validação, imagens das quais não se conhece a classe, podem ser processadas pelo BIC e seus

vetores apresentados ao classificador SVM, o qual predirá a classe de cada imagem utilizando o modelo de classificação testado.

A Figura 3.2 traz uma visão geral sobre a arquitetura do SNIF na qual pode ser observado como os seus componentes estão relacionados.

Figura 3.2: SNIF: visão geral da arquitetura.

3.2 Experimentos com o SNIF

Nesta seção são descritos os experimentos realizados para verificar a viabilidade e eficiência da abordagem proposta. De uma forma geral, os experimentos comparam o resultado da classificação fornecido por esta abordagem com a classificação manual previamente realizada ou com a classificação gerada por outros métodos propostos na literatura.

3.2.1 Ambiente de Experimentação

A seguir são apresentados os detalhes sobre o ambiente utilizado para a realização dos experimentos, incluindo informações sobre as bases de imagens de treinamento e teste utilizadas, sobre o extrator de características, parâmetros e configurações do classificador SVM. Todos os experimentos foram realizados em uma estação de trabalho Pentium 4 2.4 GHz, com 1 Gb de memória principal e um disco *IDE* com 200 Gb de espaço de armazenamento.

Bases de Imagens

As bases de imagens foram extraídas de páginas *Web* coletadas do diretório de Internet Cadê[1]. As imagens ofensivas foram extraídas das categorias relacionadas com erotismo e as imagens não-ofensivas foram coletadas de categorias como: esportes (incluindo esportes aquáticos), artes, animais e outras com fotos de pessoas vestidas. Todas as imagens usadas nas bases foram manualmente selecionadas e avaliadas para verificar se realmente continham ou não conteúdo ofensivo. Nessa avaliação, foram consideradas imagens ofensivas aquelas que exibem pessoas sem roupa ou com menos roupa do que o esperado pelas convenções da cultura local e também aquelas que, de alguma forma, exibem as partes íntimas do corpo humano. Por esse motivo, a inclusão de imagens de esportes aquáticos nas bases de treinamento e teste é uma tentativa de prever os casos onde, mesmo exibindo certas partes do corpo devido aos trajes característicos destas atividades, as imagens não possuem caráter ofensivo.

Nos experimentos do SNIF, a base de treinamento foi composta de 2.000 imagens, sendo 1.000 ofensivas e 1.000 não-ofensivas, e a base de teste foi composta de 2.135 imagens, sendo 1.000 ofensivas e 1.135 não-ofensivas.

O Extrator de Características

O extrator de característica do SNIF, especificamente a característica de cor, é baseado em uma implementação existente do BIC, a qual foi adaptada para gerar ao final do processo de extração, um histograma no formato de entrada do classificador SVM, no qual são inseridos apenas valores diferentes de zero e seus índices de posição no vetor, como comentado na Seção 3.1.1.

[1] http://www.cade.com.br

O Classificador SVM

O classificador SVM foi implementado usando a LIBSVM [6], uma biblioteca bastante utilizada, composta de um conjunto de ferramentas de grande utilidade para resolver problemas de classificação e regressão. A LIBSVM possui várias opções de *kernel*, porém o *kernel* selecionado foi do tipo RBF. Em [11] são apresentados alguns argumentos para considerar este tipo de *kernel* a escolha mais indicada, como por exemplo a menor complexidade dos cálculos devido ao número menor de parâmetros requeridos. Apenas dois parâmetros são necessários, C, o parâmetro de regularização e γ, um parâmetro interno da função *kernel*. Durante a fase de treinamento, na seleção do modelo foi aplicada a técnica de validação cruzada. Nesta técnica, o conjunto de treinamento é dividido em subconjuntos de igual tamanho e cada um é testado usando o classificador treinado com todos os outros subconjuntos. Para cada divisão adotada, foram realizados experimentos com diferentes valores para os parâmetros, variando C e γ. Ao final do processo, os melhores resultados encontrados para C e γ são utilizados para treinar o classificador utilizando a base de treinamento completa.

3.2.2 Resultados

Um primeiro experimento foi realizado com o objetivo de verificar como a quantização de cores afeta o resultado da classificação e identificar qual a quantização mais apropriada. Foram realizados testes com diferentes quantizações de cor (16, 32, 64, 128 e 256 cores), considerando os histogramas dos *pixels* de borda e de interior. Para cada quantização de cor, o classificador foi treinado e executado e, os resultados são apresentados na Tabela 3.1. Como pode ser observado, os melhores resultados foram obtidos usando 128 cores. É interessante notar que os resultados são piores ao se utilizar mais que 128 cores. Isso ocorre provavelmente pelo fato do excesso de informação gerar ruído no treinamento do classificador, ou seja, a quantidade extra de informção, que deveria auxiliar na diferenciação dos exemplos das classes, faz com que a semelhança entre os mesmos se torne maior, dificultando a classificação.

Em outro experimento, foi investigada a possibilidade de descartar parte da informação de cor, reduzindo o espaço em memória requerido para classificar as imagens e mantendo aceitável a medida F geral do classificador. Com base em observações empíricas, foi percebido que a maioria dos *pixels* cor de pele estão localizados no conjunto de *pixels* de interior, sendo assim, descartando

Cores	Especificidade	Sensibilidade	Medida F Geral
16	83.8%	94.4%	88.8%
32	84.7%	96.3%	90.1%
64	89.8%	96.4%	93.2%
128	91.0%	97.6%	94.4%
256	85.6%	97.5%	91.1%

Tabela 3.1: Resultados para classificação de imagens usando diferentes quantizações de cor.

o histograma correspondente aos *pixels* de borda, os resultados provavelmente permaneceriam estáveis.

Para confirmar esta hipótese, o classificador de *pixels* do BIC foi modificado para considerar um limiar de proximidade entre cores, ao invés de um casamento exato como originalmente proposto em [16]. Como resultado, um *pixel* é classificado como de borda se um dos seus quatro vizinhos é de uma cor fora deste limiar. Quanto maior o valor do limiar, menos *pixels* serão classificados como de borda. Diferentes valores de limiar (10, 20, 30) foram experimentados, mostrando que a sensibilidade aumenta a medida que o limiar aumenta. Isso confirma que todos os *pixels* de borda identificados pelo algoritmo padrão do BIC podem ser descartados. Como apenas o histograma dos *pixels* de interior é utilizado, apenas metade do espaço de armazenamento original é requerido. Esta característica é útil para reduzir o tempo de computação do modelo, o qual depende do número de dimensões dos vetores, e também reduz os requisitos de memória para a classificação de imagens, já que os vetores são representados com a metade do tamanho original.

Foram realizados testes com 32 e 128 cores, considerando apenas o histograma dos *pixels* de interior. Os resultados são apresentados na Tabela 3.2, onde a letra "i"ao lado do número de quantização significa que apenas os *pixels* de interior foram utilizados. Ao comparar estes resultados com os apresentados na Tabela 3.1, pode ser observado que houve uma melhoria na sensibilidade. Comparando os resultados das quantizações 128i com a de 64 cores, a melhoria é de 1.6% na sensibilidade, usando o mesmo espaço de armazenamento. Em algumas aplicações é importante obter altos valores de sensibilidade, pois isso reduz o número de imagens ofensivas apresentadas para o usuário como não-ofensivas. O inconveniente de usar apenas os *pixels* de interior foi a redução do valor de especificidade, o que significa que um número maior de imagens não-ofensivas deixam de ser apresentadas para o usuário, pois são incorretamente classificadas como ofensivas. Sendo assim, a decisão sobre o uso apenas dos *pixels* de interior depende da

aplicação dada ao método. A Tabela 3.2 também mostra que a sensibilidade para 128i foi um pouco melhor que a sensibilidade para 128, o que significa que, usando apenas os *pixels* de interior podem ser obtidos valores mais altos de revocação para as imagens ofensivas.

Cores	Especificidade	Sensibilidade	Medida F Geral
32i	87.0%	94.9%	91.0%
128i	87.4%	98.0%	92.5%

Tabela 3.2: Classificação de imagens usando apenas os *pixels* de interior.

No próximo experimento, foram testadas diferentes proporções entre as classes (ofensivas e não-ofensivas) de imagens na base de treinamento. O objetivo deste experimento foi verificar como a eficácia da classificação está relacionada com a proporção de exemplos positivos (imagens ofensivas) e negativos (imagens não-ofensivas) presentes na base de treinamento. A quantização 128i foi utilizada e três diferentes configurações foram testadas com as respectivas proporções de imagens ofensivas e não-ofensivas: (75%-25%), (50%-50%) e (25%-75%).

Os resultados na Tabela 3.3 mostram que usando relativamente poucos exemplos positivos (imagens ofensivas) é possível obter bons resultados. A suposição é que isso ocorre devido ao fato de que a classe de imagens ofensivas é bem definida, já a classe de imagens não-ofensivas é um agrupamento que inclui tipos muito diferentes de imagens, que conseqüentemente precisa de um número maior e diverso de imagens para ser definida.

Proporções	Especificidade	Sensibilidade	Medida F Geral
(75%-25%)	81.67%	98.6%	86.5%
(50%-50%)	87.58%	98.0%	92.5%
(25%-75%)	87.49%	97.8%	91.8%

Tabela 3.3: Especificidade, sensibilidade e medida F geral atingidas com diferentes proporções de imagens positivas e negativas na base de treinamento.

Em relação ao tempo de processamento, o SNIF é capaz de processar aproximadamente 16 imagens por segundo. A maior parcela do tempo total de processamento é consumida na extração das características de conteúdo da imagem. O tempo gasto pelo classificador SVM é mínimo se comparado com o tempo de extração. Levando em consideração apenas o tempo de classificação, após a extração das características, o classificador SVM é capaz de processar 1359 imagens por segundo.

Capítulo 4

Detectando Imagens Ofensivas Usando Texto

Uma alternativa para detectar imagens ofensivas na *Web* consiste na utilização de evidências textuais, uma vez que as imagens exibidas nas páginas *web* geralmente vêm acompanhadas de algum tipo de informação textual que pode fornecer indicações sobre o assunto tratado nas mesmas. A abordagem baseada em texto apresentada neste trabalho combina um extrator dessas evidências de texto presentes nas páginas HTML, com um classificador baseado em SVM. A seguir, são apresentados detalhes sobre os componentes da abordagem baseada em texto e seu processo de classificação.

4.1 Evidências Textuais

Páginas *Web* possuem uma variedade de informações que podem ser utilizadas como evidências textuais na classificação de imagens. De fato, as paginas *Web* não só possuem informação textual, como também existe uma estrutura hierárquica que permite a obtenção de alguns meta dados. Dentre as informações textuais que podem ser obtidas a partir de páginas *Web*, têm-se:

- O texto completo da página;
- Meta dados encontradas nas *tags* `<TITLE>` e `<META>`, tais como título, autor, descrição, palavras-chave, entre outras informações sobre a página;
- Informações de apontadores (*links*), encontradas nas *tags* `<A>` e `</A>`;

- Informações sobre as imagens, tais como as encontradas na *tag* `<IMG>`: nome do arquivo da imagem, descrição encontrada no atributo `ALT`;
- Passagens de texto compostas por termos encontrados próximos da localização das imagens.

A escolha de que evidências de texto melhor descrevem uma imagem não é uma tarefa simples. Coelho et al [7], por exemplo, realizou vários experimentos até encontrar um conjunto de evidências que melhor pudessem descrever uma imagem, como descrito na Seção 2.2.

4.2 Abordagem Baseada em Texto

A abordagem aqui proposta tem como base a utilização de evidências textuais para detectar imagens ofensivas. A seguir, são apresentados detalhes sobre os componentes da abordagem baseada em texto e seu processo de classificação.

4.2.1 Extração de Evidências Textuais

Segundo Coelho et al [7], que realizou a combinação de múltiplas fontes de evidências de texto para a busca de imagens na *Web*, as evidências de texto que melhor descrevem os objetos presentes em uma imagem são: o título da página *Web*, o nome da imagem sem a extensão, o conteúdo do parâmetro `ALT` da *tag* `<IMG>` e a passagem de texto composta pelos 20 termos anteriores e 20 termos posteriores à ocorrência da *tag* `<IMG>` na página.

Esta mesma combinação de evidências foi utilizada aqui como o conjunto de evidências textuais que descrevem cada imagem encontrada em uma página *Web*. O processo de extração dessas caracteríscticas é realizado por um *parser* que, ao final do processo de extração, gera um conjunto de evidências textuais para cada imagem presente em uma página *Web*. Na verdade, cada imagem passa a ser representada por um conjunto de termos, a partir do qual é composto o vetor de características a ser utilizado no classificado SVM.

Para a criação dos vetores de características, os conjuntos de evidências textuais são pré-processados por um indexador, com a intenção de compor um dicionário com os T termos diferentes presentes nestes conjuntos. Esse número de termos define o número de posições dos vetores de características que representam as imagens no classificador. Cada vetor é composto

de T posições, onde cada posição corresponde a um termo do dicionário e contém o número de ocorrências deste termo no conjunto de evidências textuais correspondente à uma imagem. Na Figura 4.1 temos alguns exemplos de termos associados às imagens ofensivas e não-ofensivas.

Figura 4.1: Exemplos de termos: (a) termos associados à imagens ofensivas e (b) termos associados à imagens não-ofensivas.

4.2.2 Treinamento Usando Texto

Da mesma forma como ocorre na abordagem baseada em cor, no treinamento da abordagem baseada em texto são necessários exemplos positivos e negativos para alimentar o classificador SVM. Neste caso, são fornecidos para o classificador exemplos positivos e negativos de conjuntos de evidências textuais. A definição de tais exemplos é feita, assim como na abordagem baseada em cor, de acordo com o conteúdo visualizado nas imagens.

Sendo assim, se uma imagem é visualmente ofensiva, o conjunto de evidências textuais, relacionado a essa imagem, é definido como um exemplo positivo, ou seja, ofensivo. Caso contrário, se uma imagem é visualmente não-ofensiva, o seu conjunto de evidências textuais é também dito não-ofensivo. A avaliação manual do texto relacionado a cada imagem se mostrou impraticável, devido ao tempo e esforço necessários a sua realização. Além disso, o julgamento sobre o caráter ofensivo de um texto é algo muito subjetivo, o que poderia causar a presença de muitos exemplos incorretos. De fato, os resultados obtidos provam que a premissa da associação de termos textuais com a imagem procede e mostram que é possível utilizar estes termos para a classificação das imagens.

Uma vez que exemplos positivos e negativos foram definidos, são gerados os vetores de características. Para cada imagem, ou seja, conjunto de evidências textuais, é gerado um vetor de características com base nos termos presente no conjunto. Os vetores de características são utilizados no treinamento do classificador SVM para gerar o modelo de classificação.

4.2.3 Teste e Classificação Usando Texto

Para a validação do modelo gerado na fase anterior, novos exemplos de conjuntos de evidências textuais são fornecidos ao classificador. Estes exemplos também são pré-classificados manualmente, como ofensivos ou não-ofensivos, através da avaliação do conteúdo das imagens as quais estão relacionados. Novamente, após a seleção dos exemplos, deve-se gerar os vetores de características, que são a informação de entrada do classificador SVM. É importante ressaltar, que ao gerar os vetores de características para esses conjuntos de evidências textuais, assim como também para outros novos conjuntos, cuja classe se deseje saber, somente os termos adicionados ao dicionário durante a fase de treinamento serão contabilizados quanto à sua freqüência, os demais serão desconsiderados.

A avaliação dos resultados produzidos pelo classificador SVM é feita com base na comparação entre a classificação manual e a produzida pelo classificador SVM. Após a validação do modelo, uma vez que o classificador tenha produzido um percentual aceitável de acertos, o classificador está pronto para predizer a classe de novas imagens, utilizando evidências textuais extraídas da mesma forma como nas fases de treinamento e teste. Detalhes sobre as métricas utilizadas na avaliação do classificador são apresentados no Capítulo 5 - Discussão.

A Figura 4.2 mostra de forma geral, como estão relacionados os componentes da abordagem baseada em texto no processo de detecção das imagens ofensivas.

4.3 Experimentos

Nesta seção são descritos os experimentos realizados para verificar a viabilidade e eficiência da abordagem proposta. De uma forma geral, os experimentos comparam o resultado da classificação fornecido por esta abordagem com a classificação manual previamente realizada ou com a classificação obtida pela aplicação de outros métodos propostos na literatura.

4.3.1 Ambiente de Experimentação

Nesta seção são apresentados detalhes sobre o ambiente utilizado para a realização dos experimentos, incluindo informações sobre as bases de imagens de treinamento e teste utilizadas, o extrator de características, parâmetros e configurações do classificador SVM. A estação de trabalho utilizada foi a mesma descrita na Seção 3.2.

Figura 4.2: Abordagem Baseada em Texto: visão geral da arquitetura.

Bases de Imagens

As bases de imagens, utilizadas nos experimentos aqui apresentados, foram compostas de forma semelhante às bases utilizadas nos experimentos da abordagem baseada em cor (Capítulo 3), ou seja, foram extraídas de páginas *Web* coletadas do diretório de Internet Cadê[1]. Nos experimentos da abordagem baseada em texto, a base de imagens foi composta de 4.406 imagens, sendo 2.000 imagens (1.000 ofensivas e 1.000 não-ofensivas) usadas no treinamento e 2.406 imagens (1.185 ofensivas e 1221 não-ofensivas) usadas no teste.

Nos experimentos da abordagem baseada em cor, a variedade de tipos de imagens era o único

[1] http://www.cade.com.br

requisito das bases utilizadas, principalmente em relação ao conjunto de imagens não-ofensivas, que compreende uma grande diversidade de imagens. Entretanto, quando se trabalha com a classificação de evidências textuais, não basta apenas garantir a variedade de tipos de imagens, mas também é necessário garantir a variedade de textos relacionados a essas imagens. Por esse motivo, na composição das bases utilizadas nos experimentos com texto, foram tomadas precauções para que nenhum *site* fosse dominante em relação ao número de imagens nas bases, prejudicando assim o treinamento e avaliação do classificador. Tentou-se normalizar o máximo possível o número de imagens por *site*, removendo algumas imagens, oriundas de *sites* com muitas imagens, e *sites* inteiros que possuíam poucas imagens.

Extrator de Características

O extrator de características da abordagem baseada em texto é dividido em duas partes principais, *parser* e indexador, que são descritos a seguir:

- *Parser*: analiza o código HTML das páginas *Web* em busca das *tags* `<IMG>`, identificando as imagens presentes na página. Para cada imagem, são extraídas e armazenadas em disco, as características descritas na Seção 4.2.1.

- Indexador: para gerar os vetores de características é utilizado o *Bow* [14], um conjunto de ferramentas para modelagem estatística, classificação e recuperação de texto. Todas as evidências textuais, extraídas pelo *parser*, são indexadas, sendo identificados os diferentes termos que compõe estas evidências e construído um dicionário de termos. Para cada termo do dicionário é atribuído um índice numérico, o qual será utilizado como índice posicional na geração dos vetores de características. O dicionário de termos é criado uma única vez, durante a fase de treinamento do classificador. Nas fases seguintes, de teste e classificação de novas imagens, o dicionário utilizado é o mesmo gerado na fase de treinamento. Uma vez que o dicionário de termos foi criado, é possível fornecer um conjunto de evidências textuais de uma imagem e obter como resposta o seu vetor de características, contendo o índice e o respectivo número de ocorrências de cada termo do dicionário no conjunto de evidências textuais informado.

Classificador SVM

No classificador SVM, também implementado usando a LIBSVM [6], foram testadas algumas configurações de *kernel*, tais como linear, polinomial e funções de base radial (RBF). Resultados preliminares obtidos com tais testes mostram que o *kernel* RBF é a melhor escolha, inclusive no caso da abordagem baseada em texto.

Também nesta abordagem, assim como na abordagem baseada em cor, foi utilizada a técnica de validação cruzada para auxiliar na escolha dos melhores valores para os parâmetros C e γ, necessários ao processo de classificação quando se utiliza o *kernel* RBF.

4.3.2 Resultados

Os experimentos utilizando texto foram realizados com as bases de imagens descritas na Seção 4.3.1, as quais são compostas por imagens e conjuntos de evidências textuais relacionados a essas imagens. Vale lembrar que esses conjuntos de evidências textuais foram criados pelo extrator de características, também descrito na Seção 4.3.1, a partir das páginas *Web* das quais as imagens foram extraídas.

Após a criação dos vetores de características pelo extrator de características, o classificador SVM foi treinado. Em seguida, a base de teste foi submetida ao classificador, obtendo os seguintes resultados:

- 98,82% de sensibilidade (revocação de imagens ofensivas);
- 99,02% de especificidade (revocação de imagens não-ofensivas);
- 98,92% de medida F geral.

Nota-se que o resultado da detecção de imagens ofensivas usando evidências de texto mostrou-se muito eficiente, com excelentes valores de sensibilidade, especificidade e consequentemente de medida F geral, o que significa que a abordagem é capaz de classificar corretamente tanto as imagens ofensivas quanto as não-ofensivas.

Capítulo 5

Discussão

Neste capítulo são discutidos os resultados obtidos nos experimentos das abordagens baseada em cor e baseada em texto. Além disso, é apresentado um estudo comparativo entre as abordagens aqui propostas e o método WIPE, proposto por Wang et al [17], devido ser a única abordagem a disponibilizar seu algoritmo para comparações.

5.1 Abordagem Baseada em Cor

Em relação aos experimentos com o SNIF, é importante ressaltar que os valores de sensibilidade obtidos são sempre expressivos, o que é uma característica interessante para este tipo de classificador, pois significa que o mesmo tem menor probabilidade de gerar falsos negativos em relação à ofensividade, ou seja, menos imagens ofensivas serão classificadas de forma incorreta como não-ofensivas.

A utilização de um número menor de cores na quantização, embora cause uma redução no valor da medida F geral, não causa uma redução acentuada nos valores de sensibilidade (revocação de imagens ofensivas), como foi mostrado na Tabela 3.1. Sendo assim, se a aplicação for um filtro de imagens pornográficas, onde a maior ênfase é não permitir que imagens ofensivas sejam exibidas, a utilização de um número menor de cores quantizadas pode ser uma opção para reduzir espaço de armazenamento e tempo de treinamento do classificador, já que os vetores de características teriam um número de posições neste caso.

Uma outra forma de obter economia de espaço e tempo, consiste em descartar os *pixels* de borda. Na Tabela 3.2, pode ser observado que ao utilizar apenas os *pixels* de interior, houve

Método	Especificidade	Sensibilidade	Medida F Geral
Forsyth e Fleck	96.0%	43.0%	-
Wang et al	91.0%	96.0%	-
Arentz e Olstad	88.0%	95.0%	89.4%
Zhu et al	-	-	94.6%
SNIF-128	91.5%	97.6%	94.4%
SNIF-128i	87.6%	98.0%	92.5%

Tabela 5.1: Comparação entre o SNIF e abordagens propostas previamente na literatura.

uma melhoria no valor de sensibilidade, com uma pequena redução no valor da medida F geral. Isso mostra que é perfeitamente aceitável descartar os *pixels* de borda, sem maiores prejuízos a qualidade dos resultados, dependendo da aplicação. Quanto a proporção de exemplos positivos e negativos utilizados no treinamento, os resultados apresentados na Tabela 3.3 indicam que a proporção de imagens ofensivas não tem um impacto significativo no valor de sensibilidade. Porém, ao utilizar uma proporção menor de imagens não-ofensivas, o valor de especificidade (revocação de imagens não-ofensivas) apresenta uma redução considerável, o que é perfeitamente compreensível, uma vez que a classe de imagens não-ofensivas é um grande aglomerado de outras possíveis classes, com uma grande diversidade de tipos de objetos presentes nas imagens.

Os resultados obtidos pelo SNIF são compatíveis com os resultados de trabalhos relacionados apresentados na Seção 2.2, como mostra a Tabela 5.1. Nesta tabela, estão presentes dois resultados do SNIF, um considerando os histogramas de borda e interior (SNIF-128) e outro considerando apenas o histograma de *pixels* de interior (SNIF-128i). Nesta mesma tabela, são apresentados também resultados de abordagens da literatura, propostas por Forsyth and Fleck [9], Wang et al [17], Arentz and Olstad [1] e Zhu et al [20], todos estes resultados foram baseados nos valores publicados nos respectivos artigos.

Na Tabela 5.1, é realizada uma comparação indireta entre o SNIF e abordagens da literatura, uma vez que são comparados resultados obtidos com bases diferentes de imagens. A natureza ofensiva das imagens dos experimentos faz com que seja difícil ter acesso às bases de imagens utilizadas em outros trabalhos. Além disso, o caráter comercial das possíveis aplicações deste tipo de classificador torna ainda mais difícil o acesso aos seus algoritmos e implementações.

Muitas imagens não-ofensivas, classificadas de forma incorreta como ofensivas (falsos positivos), são imagens de *close* de rosto. Essas imagens contém grandes áreas cor de pele, o que costuma confundir um classificador baseado em cor. Na Figura 5.1 são exibidos alguns exem-

plos de falsos positivos. No caso das imagens ofensivas, algumas das classificadas incorretamente (falsos negativos) continham muitos *pixels* de cor azul, geralmente mostrando pessoas em praias ou piscinas, como ilustrado na figura 5.2. Este fato provavelmente foi ocasionado pela presença de imagens de esportes aquáticos na base de treinamento das imagens não-ofensivas.

Figura 5.1: Falsos positivos: exemplos de imagens não-ofensivas classificadas como ofensivas.

Figura 5.2: Falsos negativos: exemplos de imagens ofensivas classificadas como não-ofensivas.

5.2 Abordagem Baseada em Texto

Embora os resultados do SNIF para sensibilidade tenham atingido valores próximos a 98% para a coleção de teste utilizada, em outras coleções, a sensibilidade apresenta variações, podendo chegar até valores próximos à 90%, dependendo do assunto, isto é, do tipo de objetos (animais, plantas, pessoas) presentes nas imagens que compõe a base de treinamento. Por outro lado, o método baseado em texto, que não depende das características de conteúdo (cor) das imagens, apresenta resultados muito similares em cada experimento. Isso ocorre mesmo nos testes iniciais, onde não foi realizada uma avaliação manual das imagens a serem processadas e aproveitou-se a classificação feita pelo próprio diretório de onde foram extraídas as imagens. Além disso,

o valores de especificidade e medida F geral foram muito próximos de 99%, o que mostra que o método baseado em texto é sem dúvida a melhor opção nos casos onde existe texto próximo à imagem a ser classificada, como acontece na *Web*. Mesmo quando existe um número pequeno de termos próximo às imagens, o classificador consegue predizer corretamente a classe da imagem. A maioria das classificações incorretas acontecem nos casos onde não há texto, ou quando além de poucos termos, estes não são representativos em relação a nenhuma das classes. A abordagem baseada em texto, além de apresentar os melhores resultados em termo de sensibilidade e especificidade, tem a vantagem de não necessitar do processamento de imagens para a extração de características, o que lhe garante um melhor desempenho. Estes resultados mostram que, ao contrário do que diz a literatura [17][1], o texto é uma fonte de evidência eficaz na detecção de imagens ofensivas, desde que utilizado de forma correta, superando os resultados das abordagens baseadas em conteúdo.

5.3 Combinação das Abordagens Baseada em Cor e Baseada em Texto

Foram realizadas algumas tentativas de encontrar uma relação entre os resultados das abordagens baseada em cor e texto para combiná-los e, assim obter um classificador mais eficaz. A idéia inicial era aplicar o SNIF nos casos onde a abordagem baseada em texto apresentasse maior probabilidade de falha, uma vez que o SNIF tem custo computacional maior que o apresentado pela abordagem baseado em texto. Entretanto, não se obteve sucesso em nenhuma das tentativas de combinação. O resultado final da combinação foi sempre inferior aos resultados obtidos individualmente por cada abordagem. Aparentemente, não existe uma relação direta entre as imagens classificadas de forma incorreta nas duas abordagens, o que torna difícil a utilização de uma abordagem para complementar a outra. Entretanto, não se pode afirmar que não seja possível realizar uma combinação das duas abordagens obtendo resultados eficazes.

Ambas as abordagens podem ser utilizadas separadamente, com resultados satisfatórios, dependendo da situação em que são aplicadas. Por exemplo, quando houver texto associado à imagem, a abordagem baseada em texto é a mais indicada, uma vez que apresenta os melhores resultados. Por outro lado, caso as imagens, a serem classificadas, não venham acompanhadas de informação textual, o SNIF pode ser utilizado também apresentando bons resultados.

Método	Especificidade	Sensibilidade	Medida F Geral
WIPE	60.2%	88.8%	74.3%
SNIF-128	89.3%	86.4%	87.9%
Texto	99.0%	98.8%	98.9%

Tabela 5.2: Comparação direta entre as abordagens aqui propostas e o método WIPE.

5.4 Comparação das Abordagens

São apresentados na Tabela 5.2 os resultados obtidos pelas abordagens aqui propostas e pelo método WIPE [17]. Para esta comparação, a base de imagens, utilizada nos experimentos da abordagem baseada em texto, foi também processada pelo SNIF e pelo método WIPE, o qual é disponibilizado como serviço na página *http://wang.ist.psu.edu/WIPE/*, desta forma, uma comparação direta pode ser realizada, já que a mesma base de imagens pode ser processadas pelos métodos a serem comparados, inclusive o WIPE. Neste serviço, é informado ao WIPE o endereço da imagem a ser classificada e como saída têm-se a classe da imagem (ofensiva ou não-ofensiva). Para obter a classificação do método WIPE, cada imagem da base é submetida à este serviço, de forma manual.

Como pode ser observado na Tabela 5.2, embora o SNIF e o método WIPE tenham apresentado valores de sensibilidade próximos, o WIPE apresenta um valor de especificidade bem inferior ao SNIF, o que significa que o WIPE tende a gerar um número maior de falsos positivos. Sendo assim, para a base de imagens utilizada, o SNIF obteve resultados melhores do que os obtidos pelo WIPE, uma vez que existe uma diferença considerável entre os valores de medida F geral obtidos. É importante lembrar que a medida F geral é uma métrica que mensura a qualidade dos resultados como um valor unificado, considerando a detecção tanto das imagens ofensivas, quanto das imagens não-ofensivas. Em relação ao desempenho, o SNIF também apresenta melhores resultados, uma vez que é capaz de processar 16 imagens por segundo enquanto o WIPE pecisa de aproximadamente 10 segundos para processar uma única imagem, como relatado em [17]. O método baseado em texto apresenta resultados expressivamente melhores do que os obtidos com os métodos baseados em conteúdo, confirmando-se assim como a opção mais indicada para aplicações na *Web*.

Capítulo 6

Conclusão e Trabalhos Futuros

Neste trabalho foram apresentadas duas abordagens para detecção automática de conteúdo ofensivo na *Web*, uma baseada no conteúdo da imagem, especificamente as cores presentes na mesma, e outra baseada nas evidências de texto associadas à imagem.

Na abordagem baseada em conteúdo, chamada *Simple Nude Image Finder* - SNIF, foram combinados um algoritmo para extração de características de cor das imagens, o BIC (*Border / Interior Pixel Classification*), e uma técnica de aprendizagem de máquina, o SVM, para compor um classificador capaz de predizer à que classe (ofensiva ou não-ofensiva) pertence uma determinada imagem.

Foram realizados experimentos quanto ao impacto da quantização na qualidade dos resultados do classificador. Verificou-se que ao utilizar uma quantização com um número menor de cores, embora o valor da medida F geral sofra uma pequena perda (Tabela 3.1), o valor de sensibilidade permanece estável. Isso significa que, dependendo da aplicação, uma quantização com um número menor de cores pode ser uma alternativa para economia de espaço de armazenamento, devido à redução de tamanho dos vetores de características. Outro fator que proporciona economia de espaço é a possibilidade de descartar informação sobre os *pixels* de borda, o que foi comprovado também através de experimentos descritos na Seção 3.2.2, com uma redução de 50% do espaço de armazenamento originalmente requerido.

Apesar de sua simplicidade, por utilizar apenas uma característica, a abordagem baseada em conteúdo apresentou bons resultados, com a vantagem de proporcionar economia de espaço de armazenamento das informações necessárias para classificar as imagens. Além disso, esta abordagem apresenta uma alta sensibilidade, sem comprometer a especificidade, isto é, a abordagem é bastante eficaz na detecção de imagens ofensivas, sem que isso aumente drasticamente

a ocorrência de falsos positivos (imagens não-ofensivas classificadas como ofensivas).

A abordagem baseada em texto também é resultado da combinação de um algoritmo de extração de características e um classificador SVM. Entretanto, nesta abordagem, são utilizadas como características das imagens, evidências textuais extraídas das páginas *Web* onde se encontram as imagens avaliadas.

A abordagem baseada em texto se mostrou a mais efetiva na detecção de imagens ofensivas da *Web*, apresentando excelentes resultados nos experimentos realizados, com valores de sensibilidade e especificidade próximos à 99%. Isso mostra que esta abordagem é capaz de classificar com altíssimo grau de corretude tanto imagens ofensivas quanto imagens não-ofensivas.

As tentativas de combinação de resultados das abordagens propostas não apresentaram resultados satisfatórios, o que leva a conclusão de que a melhor alternativa é utilizá-las de forma isolada, aplicando cada abordagem de acordo com a situação. Quando houver texto associado às imagens, a abordagem baseada em texto é a mais indicada, e quando não houver texto, o SNIF é uma alternativa válida.

6.1 Trabalhos futuros

Um possível trabalho futuro é a implementação de aplicações práticas para as abordagens aqui propostas, como por exemplo, uma espécie de *plug-in* para navegadores *Web* ou um módulo adicional para servidores *proxy* com o objetivo de impedir o acesso à conteúdo ofensivo na *Web*. Nesse caso, um aspecto importante é o desempenho, o que pode indicar que a abordagem baseada em texto é uma boa opção para este tipo de aplicação, porém, nada impede que sejam utilizadas várias estações em paralelo para obter um melhor desempenho na utilização da abordagem baseada em conteúdo da imagem.

Outra possibilidade é a aplicação das duas abordagens propostas em outros domínios específicos, diferentes do domínio ofensivo. Desde que estes novos domínios também tenham características bem-definidas, que possam distinguir um conjunto de imagens, tais como a presença predominante de determinadas cores, como ocorre na abordagem baseada em cor, ou a presença de um conjunto de termos específicos, como ocorre na abordagem baseada em texto.

Também pode ser verificada a utilização de outros algoritmos para extração de características de conteúdo, que sejam igualmente simples, visando melhorar a qualidade da classificação sem perda de performance.

Referências Bibliográficas

[1] Will Archer Arentz and Bjorn Olstad. Classifying offensive sites based on image content. *Comput. Vis. Image Underst.*, 94(1-3):295–310, 2004.

[2] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[3] Ruan J. S. Belém, João M. B. Cavalcanti, Edleno Silva de Moura, and Mario A. Nascimento. Snif: A simple nude image finder. In *LA-WEB*, pages 252–258, 2005.

[4] J. Bilmes. A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models, 1997.

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

[6] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software disponível em http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[7] Tatiana Almeida Souza Coelho, Pável Pereira Calado, Lamarque Vieira Souza, Berthier Ribeiro-Neto, and Richard Muntz. Image retrieval using multiple evidence ranking. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):408–417, 2004.

[8] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[9] D. A. Forsyth and M. M. Fleck. Identifying nude pictures. In *Proceedings of the 3rd IEEE Workshop on Applications of Computer Vision (WACV '96)*, page 103. IEEE Computer Society, 1996.

[10] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (2nd Edition)*. Addison-Wesley, 2002. GON r 02:1 1.Ex.

[11] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification.

[12] Varghese Jacob, Ramayya Krishnan, Young U. Ryu, R. Chandrasekaran, and Sungchul Hong. Filtering objectionable internet content. In *ICIS '99: Proceeding of the 20th international conference on Information Systems*, pages 274–278, Atlanta, GA, USA, 1999. Association for Information Systems.

[13] D. Mazzoni, K. Wagstaff, and R. Castano. Using trained pixel classifiers to select images of iterest. In *The Interplanetary Network Progress Report*.

[14] Andrew Kachites McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. Software disponível em http://www.cs.cmu.edu/~mccallum/bow, 1996.

[15] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[16] Renato O. Stehling, Mario A. Nascimento, and Alexandre X. Falcao. A compact and efficient image retrieval approach based on border/interior pixel classification. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 102–109. ACM Press, 2002.

[17] J. Wang, J. Li, G. Wiederhold, and O. Firschein. System for screening objectionable images. *Computer Communications Journal*, 1998.

[18] James Wang, J. Li, G. Wiederhold, and O. Firschein. System for screening objectionable images using daubechies' wavelets and color histograms. In *Proceedings of the Interactive Distributed Multimedia Systems (IDMS'97)*, volume 1309, pages 20–30. Springer-Verlag LNCS, 1997.

[19] Yi-Leh Wu, Edward Y. Chang, Kwang-Ting Cheng, Chengwei Chang, Chen-Cha Hsu, Wei-Cheng Lai, and Ching-Tung Wu. Morf: A distributed multimodal information filtering system. In *IEEE Pacific Rim Conference on Multimedia*, pages 279–286, 2002.

[20] Qiang Zhu, Ching-Tung Wu, Kwang-Ting Cheng, and Yi-Leh Wu. An adaptive skin model and its application to objectionable image filtering. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 56–63. ACM Press, 2004.