USO DE DESCRITORES BINÁRIOS PARA DETECÇÃO DE PORNOGRAFIA

CARLOS ANTÔNIO CAETANO JÚNIOR

USO DE DESCRITORES BINÁRIOS PARA DETECÇÃO DE PORNOGRAFIA

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

Orientador: Arnaldo de Albuquerque Araújo
Coorientador: Silvio Jamil Ferzoli Guimarães

Belo Horizonte
Maio de 2014

Caetano Júnior, Carlos Antônio
C127u Uso de Descritores Binários para Detecção de Pornografia / Carlos Antônio Caetano Júnior. — Belo Horizonte, 2014
xxii, 74 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Arnaldo de Albuquerque Araújo
Coorientador: Silvio Jamil Ferzoli Guimarães

1. Visão Computacional. 2. Características Locais. 3. Descritores Binários. 4. BossaNova. 5. Reconhecimento Visual. 6. Pornografia. I. Orientador. II. Coorientador. III. Título.

CDU 519.6*85(043)

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

FOLHA DE APROVAÇÃO

Uso de descritores binários para detecção de pornografia

CARLOS ANTONIO CAETANO JUNIOR

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

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

PROF. SILVIO JAMIL FERZOLI GUIMARÃES - Coorientador
Instituto de Ciências Exatas e Informática - PUC/MG

PROF. EDUARDO ALVES DO VALLE JUNIOR
Departamento de Engenharia da Computação e Automação - UNICAMP

PROF. JEFERSSON ALEX DOS SANTOS
Departamento de ciência da Computação - UFMG

PROF. WILLIAM ROBSON SCHWARTZ
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 30 de maio de 2014.

Agradecimentos

A Deus por ter sempre iluminado meu caminho e guiado meus passos nessa jornada.

Ao Prof. Arnaldo de Albuquerque Araújo pela orientação e pela oportunidade concedida para fazer parte do Núcleo de Processamento Digital de Imagens (NPDI) da UFMG.

A meu coorientador, Silvio Jamil Ferzoli Guimarães, pelas discussões, sugestões, críticas e ideias fundamentais que colaboraram no desenvolvimento deste trabalho. Meus sinceros agradecimentos.

A minha coorientadora Sandra Eliza Fontes de Avila pelos conhecimentos transmitidos, amizade e conselhos. Sem a sua participação e longas discussões que tivemos este trabalho não estaria no nível que se encontra. Meus sinceros agradecimentos.

Aos meus pais, Carlos e Lucilda, pelo amor, carinho, educação, apoio e pelo exemplo de pai e mãe que representam.

A minha irmã, Narjara, e minha sobrinha, Júlia, pelo carinho e apoio.

A minha namorada Ana Paula, por todo seu amor, companheirismo, carinho e por sempre me apoiar em todas as minhas decisões.

A minha família pelo afeto e pela compreensão dos momentos ausentes.

Aos colegas do NPDI: Alberto, Andrey, Bruno, Cássio, Guilherme, Henrique, Kleber, Lilian, Suellen, Tiago e Virgínia. Muito obrigado pelo apoio e pelos momentos de descontração.

Aos professores do DCC, pelo ensino e conhecimento adquirido.

Ao CNPq, CAPES, FAPEMIG e projetos SmartView e InWeb pelo imprescindível apoio financeiro.

Por fim, a todos que colaboraram diretamente ou indiretamente na execução deste trabalho. Muito Obrigado!

"Ciência da Computação está tão relacionada aos computadores quanto a Astronomia aos telescópios, Biologia aos microscópios, ou Química aos tubos de ensaio. A Ciência não estuda ferramentas. Ela estuda como nós as utilizamos, e o que descobrimos com elas."

(Edsger Dijkstra)

Resumo

Com o crescimento da quantidade de conteúdos inapropriados na internet, como pornografia, surge uma necessidade de detectar e filtrar tal tipo de material. O motivo disto é dado pelo fato de que esse tipo de conteúdo é frequentemente proibido em certos ambientes (como, escolas e locais de trabalho) ou para certos públicos (crianças).

Nos últimos anos, diversos trabalhos da literatura têm tido como foco principal detectar imagens e vídeos pornográficos baseados em conteúdo visual, principalmente a detecção de cor de pele. Apesar de apresentarem bons resultados, essas abordagens geralmente têm como desvantagem uma alta taxa de falsos positivos, pois nem todas imagens com grandes áreas de exposição de pele são necessariamente pornográficas, como imagens com pessoas usando roupas de banho, ou imagens relacionadas a esportes. Abordagens baseadas em características locais, em conjunto com modelos *Bag-of-Words* (BoW), têm sido aplicadas com sucesso em tarefas de reconhecimento visual no contexto de detecção de pornografia. Apesar dos métodos existentes produzirem resultados promissores no contexto de detecção de pornografia, estes fazem uso de descritores de características locais que necessitam de um alto tempo computacional de processamento, além de gerarem vetores de alta dimensionalidade.

Neste trabalho, é proposta uma abordagem simples, eficaz e eficiente para o problema de reconhecimento visual no contexto de detecção de pornografia. O método é baseado na extração das características locais utilizando descritores binários, uma alternativa de baixa complexidade, em conjunto com a recente representação intermediária BossaNova, uma extensão do modelo BoW, que preserva, de uma maneira mais rica, a informação visual. Além disso, é proposto de um descritor de vídeo baseado na combinação de representações intermediárias. Os resultados obtidos validaram a abordagem proposta e o descritor de vídeo apresentando resultados com qualidade superior em relação às demais abordagens encontradas na literatura.

Palavras-chave: Visão Computacional, Características Locais, Descritores Binários, BossaNova, Reconhecimento Visual, Pornografia.

Abstract

With the growing of the amount of inappropriate content on the Internet, such as pornography, it arises the need to detect and filter such material. The reason for this is given by the fact that such content is often prohibited in certain environments (e.g., schools and workplaces) or for certain publics (e.g., children).

In recent years, many works of the literature have been mainly focused on detecting pornographic images and videos based on visual content, particularly on the detection of skin color. Although these approaches provide good results, they generally have the disadvantage of a high false positive rate, since not all images with large areas of skin exposure are necessarily pornographic images, such as people using swimsuits or images related to sports. Local feature based approaches, with Bag-of-Words models (BoW), have been successfully applied to visual recognition tasks in the context of pornography detection. Despite existing methods provide promising results in the context of detection of pornography, they use local features descriptors that require a high computational processing time, and generate high-dimensional vectors.

In this work, we propose a simple, effective and efficient approach to the problem of visual recognition in the context of pornography detection. The method is based on local feature extraction using binary descriptors, a low-complexity alternative, in conjunction with the recent mid-level representation BossaNova, a BoW model extension that preserves more richly the visual information. Moreover, we propose a video descriptor based on the combination of mid-level representations. The results validated the proposed approach and the video descriptor by presenting results with superior quality compared to other approaches in the literature.

Keywords: Computer Vision, Local Features, Binary Descriptors, BossaNova, Visual Recognition, Pornography.

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

Acc	Acurácia
AGAST	*Adaptive and Generic Accelerated Segment Test*
AP	*Average Precision*
BoF	*Bag-of-Features*
BoW	*Bag-of-Words*
BRIEF	*Binary Robust Independent Elementary Features*
BRISK	*Binary Robust Invariant Scalable Keypoints*
CenSurE	*Center-Surround Extrema*
COR	*Correlation Coefficient*
DFM	*Double-Fault Measure*
DM	*Disagreement Measure*
DoG	*Differences of Gaussians*
FAST	*Fast Accelerated Segment Test*
FREAK	*Fast Retina Keypoint*
GLOH	*Gradient Location and Orientation Histogram*
GMM	*Gaussian Mixture Model*
HOG	*Histograms of Oriented Gradients*
HSV	*Hue, Saturation, Value*
IA	*Interrater Agreement k*

LBP *Local Binary Pattern*

LoG *Laplacian of Gaussian*

mAP *mean Average Precision*

MFCC *Mel-Frequency Cepstral Coefficients*

ORB *Oriented Fast and Rotated BRIEF*

PCA *Principal Component Analysis*

QSTAT *Q-Statistic*

RGB *Red, Green, Blue*

SIFT *Scale Invariant Feature Transform*

STIP *Space-Time Interest Points*

SURF *Speeded Up Robust Feature*

SVM *Support Vector Machines*

VD *Video Descriptor*

VOC *Visual Object Classes*

Sumário

Capítulo 1

Introdução

Com o rápido crescimento da quantidade de imagens e vídeos disponíveis publicamente na Internet, surge uma necessidade de reconhecimento em relação ao conteúdo destes materiais. Além da necessidade óbvia de métodos relacionados a pesquisas de imagens e vídeos, também é importante realizar o reconhecimento de conteúdos que podem ser considerados indesejados ou ofensivos, a fim de ser capaz de filtrar estes materiais.

O maior grupo de imagens e vídeos disponíveis na Internet que as pessoas podem considerar ofensivo é referente a materiais pornográficos. Um relatório publicado no *site* de tecnologia ExtremeTech[1] sugere que 30% de todo o tráfego da Internet está associado a pornografia. Eles chegaram a este número estimando o tráfego que um *site* popular, de conteúdo pornográfico, gera a cada dia e multiplicaram pelos vários outros *sites* pornográficos, de dimensão semelhante, encontrados na Internet. Além disso, segundo o relatório da ExtremeTech, o Xvideos[2], que é o maior *site* fornecedor deste tipo de conteúdo, recebe três vezes mais *pageviews* do que grandes *sites* de notícias, como CNN ou ESPN (cerca de 4,4 bilhões de *pageviews* por mês) e o tempo médio gasto neste *site* chega a ser cinco vezes maior do que nos *sites* de notícias.

Detectar e filtrar conteúdo visual pornográfico proveniente da Internet é uma preocupação em vários ambientes, desde casas com crianças e escolas até locais de trabalho. *Tags* textuais vinculadas às imagens e vídeos claramente não são suficientes, uma vez que conteúdos impróprios podem maliciosamente estar anexados a textos aparentemente inocentes [Lopes et al., 2009b]. Uma situação típica seria, por exemplo, utilizar palavras-chave de busca comumente usadas por crianças anexadas em *sites* com conteúdo pornográfico. Porém, adultos também podem não querer ser expostos a tal conteúdo, como por exemplo, proveniente de resultados que recebem a partir de

[1] `http://www.extremetech.com/computing/123929-just-how-big-are-porn-sites`
[2] `http://www.xvideos.com`

máquinas de busca disponíveis na web.

Uma definição comumente utilizada pela literatura é que a pornografia pode ser considerada como "qualquer material sexualmente explícito com o objectivo de excitação sexual ou fantasia" [Short et al., 2012]. Porém, esta definição pode levar a diversos desafios quando se está tentando detectar conteúdo pornográfico, como por exemplo a definição de qual o limite de "explícito" deve ser considerado para que algo seja considerado como material pornográfico. Alguns autores lidam com essa questão dividindo o material em diversas classes [Deselaers et al., 2008], complicando ainda mais a tarefa de classificação. Por outro lado, alguns autores optam por tratar do assunto utilizando uma avaliação conceitualmente simples através da atribuição de apenas duas classes (pornográfica e não-pornográfica) [Valle et al., 2011; Avila et al., 2013].

Nos últimos anos, diversos trabalhos da literatura têm tido como foco principal detectar imagens e vídeos pornográficos baseados em conteúdo visual em vez de informações textuais [Forsyth & Fleck, 1996, 1997, 1999; Jones & Rehg, 2002; Zheng et al., 2004; Rowley et al., 2006; Hu et al., 2007; Deselaers et al., 2008; Lopes et al., 2009a; Ulges & Stahl, 2011; Valle et al., 2011; Steel, 2012; Avila et al., 2013; Yu & Han, 2014]. A maioria destes trabalhos é feita baseada em abordagens de detecção de cor de pele. Isto se dá pelo fato de que a propriedade mais óbvia em imagens pornográficas é uma grande fração de *pixels* que apresentam cores relacionadas a pele [Ries & Lienhart, 2014]. Apesar disso, essas abordagens geralmente têm como desvantagem uma alta taxa de falsos positivos, pois nem todas imagens com grandes áreas de exposição de pele são necessariamente pornográficas (imagens com pessoas usando roupas de banho, ou imagens relacionadas a esportes). Também, outro problema a ser observado é que, imagens em escalas de cinza não podem ser classificadas usando características relacionadas a cor.

Segundo Lopes et al. [2009b], a tarefa de detecção de pornografia pode ser interpretada como uma tarefa de reconhecimento visual no contexto de detecção de objetos. Abordagens baseadas em características locais, em conjunto com modelos *Bag-of-Words* (BoW), têm sido aplicadas com sucesso em tarefas de classificação em reconhecimento visual [Agarwal et al., 2004; Yang et al., 2007]. Em tais abordagens, as imagens são representadas como histogramas construídos a partir de um conjunto de características visuais. Nenhum modelo explícito do objeto é necessário e a variabilidade de exemplos (relacionados a forma, escala ou iluminação) é tratada por um conjunto de treinamento que abrange essa variabilidade. Estas características tornam o uso de abordagens com o modelo BoW adequadas para o contexto de detecção de pornografia.

Apesar dos métodos existentes, baseados em características locais, produzirem

resultados promissores no contexto de detecção de pornografia, estes fazem uso de descritores de características locais que necessitam de um alto tempo computacional de processamento, além de gerarem vetores de alta dimensionalidade compostos por valores reais. Por exemplo, Avila et al. [2013] fazem uso do descritor de características HueSIFT [Van de Sande et al., 2010], uma variação do descritor SIFT (*Scale Invariant Feature Transform*) que inclui informação de cor, que leva em média um tempo de 2,5 segundos para extrair, de maneira densa, as características locais de uma imagem, gerando um vetor de características composto por 165 valores de ponto flutuante. De fato, isso ainda não é rápido o suficiente para aplicações que necessitam de um curto tempo de resposta. Além do mais, a comparação entre duas características extraídas gastaria mais tempo computacional, devido à alta dimensionalidade.

Neste trabalho, é proposta uma abordagem simples, eficaz e eficiente para o problema de reconhecimento visual no contexto de detecção de pornografia em vídeos. O método é baseado na extração das características locais utilizando descritores binários, uma alternativa de baixa complexidade, em conjunto com a recente representação intermediária BossaNova, uma extensão do modelo BoW que preserva, de maneira mais rica, a informação visual. Além disso, é proposto de um descritor de vídeo baseado na combinação de representações intermediárias. Os resultados alcançados foram comparados com abordagens encontradas na literatura e com o software PornSeer Pro[3], um sistema de detecção de pornografia da indústria.

1.1 Motivação

O presente trabalho é motivado pelo crescimento exponencial de conteúdos que podem ser considerados inapropriados na Internet, como pornografia. Com o aumento no volume deste tipo de conteúdo, cresce também a necessidade de desenvolvimento de técnicas para detectar e filtrar tal tipo de material. Uma classificação precisa sobre este tipo conteúdo ajudará para decidir se um determinado usuário final pode ou não ter acesso a este tipo de material.

Além disso, apesar dos recentes avanços na área de reconhecimento visual no contexto de detecção de pornografia, essa tarefa ainda se constitui um desafio visto que a utilização de técnicas tradicionais, relacionadas a cor e detecção de pele, nem sempre são adequadas devido a alta taxa de falsos positivos que podem gerar. Ademais, técnicas relacionadas a características locais das imagens/vídeos podem ser exploradas. Portanto, apesar dessa estratégia ter demonstrado resultados promissores no contexto

[3] `http://www.yangsky.com/products/dshowseer/porndetection/PornSeePro.htm`

de detecção de pornografia, estas fazem uso de descritores de características locais que necessitam de um alto tempo computacional de processamento, gerando vetores de alta dimensionalidade compostos por valores reais. Desta maneira, como uma alternativa de baixa complexidade, descritores binários podem ser aplicados por gerarem resultados similares, ou até mesmo superiores, quando comparados a descritores não-binários do estado da arte.

1.2 Objetivos

1.2.1 Objetivo Geral

Desenvolver uma abordagem eficaz e eficiente para detecção de pornografia, que gere uma classificação precisa, possibilitando identificar se um determinado vídeo contém conteúdo pornográfico.

1.2.2 Objetivos Específicos

1. Validar os descritores binários para tarefas de classificação em reconhecimento visual;

2. Comparar a performance dos descritores binários para tarefas de classificação no contexto de detecção de pornografia;

3. Propor uma metodologia simples para detecção de pornografia que classifique vídeos pornográficos de uma forma eficaz e eficiente.

1.3 Contribuições

Este trabalho apresenta três contribuições principais: (i) a definição de um método simples, eficaz e eficiente para detecção de pornografia; (ii) a proposta de um descritor de vídeo, baseado na combinação de representações intermediárias; e (iii) um estudo sobre descritores binários em contraponto aos descritores não-binários. A metodologia proposta integrou as vantagens dos conceitos apresentados nos trabalhos de referência na área de detecção de pornografia com recentes características locais de imagens, conhecidas como descritores binários, contribuindo para o aprimoramento do estado da arte desta área de aplicação.

Também, pode ser citado como contribuição, os experimentos de validação do uso de descritores binários para tarefas de reconhecimento visual na base de dados PASCAL

VOC 2007, onde, a solução proposta apresentou resultados com qualidade superior em relação às demais abordagens encontradas na literatura que também empregaram descritores binários.

1.4 Organização do Documento

O restante desta dissertação está organizada como a seguir. O Capítulo 2 fornece os fundamentos básicos para a compreensão dos métodos e algoritmos utilizados e mencionados neste trabalho. O Capítulo 3 apresenta uma revisão bibliográfica sobre os principais métodos de reconhecimento visual no contexto de detecção de pornografia. O Capítulo 4 descreve a metodologia desenvolvida neste trabalho para detecção de pornografia. Posteriormente, no Capítulo 5, são apresentadas as bases de dados utilizadas nos experimentos, os resultados obtidos e a análise destes. Finalmente, uma conclusão e discussão sobre trabalhos futuros são apresentados no Capítulo 6.

Capítulo 2

Fundamentação Teórica

Um dos maiores desafios da área de visão computacional é permitir que computadores sejam capazes de compreender os dados contidos em imagens e vídeos, ou seja, tarefas de reconhecimento visual. Devido à imensa disponibilidade de imagens e vídeos, fornecidos por câmeras digitais e disponibilizados na Internet, torna-se necessário o uso de sistemas automatizados inteligentes o suficiente para categorizar, monitorar e filtrar (caso seja necessário) objetos, lugares ou até mesmo pessoas e suas ações. Por este motivo, diversos trabalhos têm surgido com foco em abordagens que envolvem o reconhecimento visual [Sivic & Zisserman, 2003; Csurka et al., 2004; Lazebnik et al., 2006; Gosselin et al., 2008; Avila et al., 2012].

Segundo Chatfield et al. [2011], a abordagem de reconhecimento visual mais utilizada na literatura é composta por três etapas distintas: (i) extração de características locais da imagem; (ii) codificação das características locais em uma representação intermediária (*mid-level*); e (iii) classificação da representação intermediária, geralmente, baseada em técnicas de aprendizado de máquina. Normalmente, as características locais extraídas tendem a ser invariantes sobre algumas transformações causadas por mudanças de câmera, como variações de rotação, escala, iluminação, dentre outras. Para lidar com essas transformações, o mais comum é extrair essas características locais utilizando os descritores SIFT (*Scale Invariant Feature Transform*) [Lowe, 2004] e SURF (*Speeded Up Robust Feature*) [Bay et al., 2008]. Em relação à representação intermediária, o modelo *Bag-of-Words* (BoW) [Sivic & Zisserman, 2003] é a abordagem mais comum utilizada para fazer a codificação das características locais extraídas das imagens. Por fim, o objetivo da classificação da representação intermediária é aprender uma função que possa atribuir rótulos (discretos) às imagens. Desta maneira, a maioria dos trabalhos de reconhecimento visual fazem uso de técnicas de aprendizado de máquina, como *Support Vector Machines* (SVM).

2.1 Descrição de Imagens Baseada em Características Locais

Segundo Tuytelaars & Mikolajczyk [2008], características locais consistem em padrões de imagem que se diferem de sua vizinhança, geralmente associados a mudanças nas propriedades da imagem (por exemplo textura e contraste). A extração de características locais é a primeira etapa a ser feita em um processo que envolva reconhecimento visual. Uma maneira de se realizar tal etapa consiste em selecionar *patches* (regiões, em português) da imagem que contenham informações relevantes, e então descrevê-las com o uso de algum descritor de características.

2.1.1 Seleção da Região a Ser Descrita

De acordo com Tuytelaars [2010], a seleção dos *patches* pode ser feita com base em dois tipos de abordagens: (i) utilizando pontos de interesse, neste caso é aplicado um algoritmo para encontrar tal região a ser descrita; ou (ii) amostragem densa, onde regiões de tamanho fixo são alocadas em uma grade de tamanho regular. A Figura 2.1 ilustra a quantidade de características locais extraídas com cada abordagem.

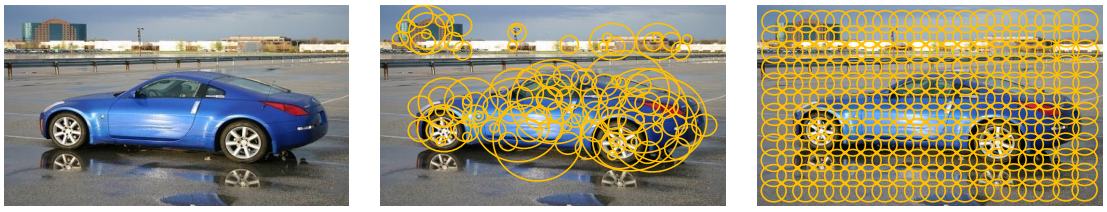


Imagem original | Pontos de Interesse | Amostragem Densa

Figura 2.1. Exemplo de características locais extraídas pelas abordagens de pontos de interesse e amostragem densa. Cada círculo amarelo representa uma característica local a ser extraída.

2.1.1.1 Pontos de Interesse

Segundo Canclini et al. [2013], os detectores de pontos de interesse encontrados na literatura podem ser divididos em dois grupos principais: (i) *corner detectors*, detectores de borda ou canto; e (ii) *blob detectors*, após a aplicação de um filtro na imagem, tentam encontrar extremos locais.

É importante ressaltar que a lista a seguir não é uma lista exaustiva de todos detectores de pontos de interesse presentes na literatura. No entanto, pode ser considerada como um grupo representativo dos detectores mais relevantes. Mais

informações sobre detectores podem ser encontradas em: [Schmid et al., 2000; Mikolajczyk & Schmid, 2005; Tuytelaars & Mikolajczyk, 2008; Gauglitz et al., 2011; Miksik & Mikolajczyk, 2012; Canclini et al., 2013].

Corner Detectors

Segundo Trajkovic & Hedley [1998], um canto pode ser definido como um determinado ponto em que existem duas direções dominantes diferentes de bordas na vizinhança deste mesmo ponto. O uso de cantos ou bordas estão entre as primeiras características de baixo nível utilizadas para análise de imagem, mais precisamente usadas para realizar *tracking* [Moravec, 1980].

Baseando-se na pesquisa de Moravec, Harris & Stephens [1988] analisaram o gradiente da imagem levando em consideração que em um canto a intensidade da imagem irá alterar grandemente em múltiplas direções, gerando então um detector que é amplamente conhecido hoje como detector de Harris. Mikolajczyk & Schmid [2001] propuseram uma abordagem para tornar o detector de Harris invariante a mudanças de escala (Harris-Laplace), combinando o detector com uma seleção de escala baseada no operador de Laplace [Lindeberg, 1998]. Uma versão mais atualizada do detector Harris-Laplace foi proposta em Mikolajczyk & Schmid [2004] promovendo mais pontos de interesse, porém com uma precisão ligeiramente inferior. Também, no mesmo trabalho, foi proposta uma extensão do detector de Harris que fosse invariante a transformações afins, intitulado Harris-Affine.

De uma maneira diferente, Trajkovic & Hedley [1998] desenvolveram um detector em que o valor do *pixel* central de uma determinada região é comparado com os valores de outros *pixels* dentro desta mesma região. Rosten & Drummond [2006] aprimoraram esta ideia com uma abordagem baseada em aprendizado de máquina para criar árvores de decisão que permitam ao detector classificar um ponto candidato com menos comparações entre *pixels*, dando origem ao detector FAST (*Fast Accelerated Segment Test*). Mair et al. [2010] apresentaram o detector AGAST (*Adaptive and Generic Accelerated Segment Test*), aumentando a performance de seu antecessor, FAST, em 50%, alterando a maneira em que as árvores de decisão são criadas.

Blob Detectors

Em vez de tentar detectar cantos, os detectores de *blobs* detectam pontos de interesse com base em extremos locais após a aplicação de filtros sobre a imagem. Normalmente, esses filtros são projetados para serem aproximações do filtro *Laplacian of Gaussian* (LoG). Em geral, pontos detectados por detectores de *blobs* tendem a ser menos precisos do que pontos detectados por detectores de canto, pois a localização

de um canto pode ser identificada por um único ponto, enquanto *blobs* só podem ser localizados através de seus limites, que são muitas vezes irregulares. Por outro lado, suas propriedades de escala e formato são mais bem definidas, pois a estimativa de escala de um canto é mal definida, como por exemplo, na interseção de arestas existe uma ampla gama de escalas. Os limites de um *blob* no entanto, mesmo que irregulares, dão uma boa estimativa do tamanho da escala do *blob* [Tuytelaars & Mikolajczyk, 2008].

Como precursor, entre os detectores de *blobs*, o detector de Hessian [Beaudet, 1978] procura por locais da imagem que apresentem mudanças em duas direções ortogonais usando o determinante de uma matriz Hessiana. Como no caso do detector de Harris, com o objetivo de gerar mais pontos de interesse, obter invariância a transformações afins e invariância a mudanças de escala, Mikolajczyk & Schmid [2004] apresentaram duas extensões do descritor: Hessian-Laplace e Hessian-Affine.

Lowe [2004] apresentou um detector de pontos de interesse invariante a rotação e escala. Conhecido como detector SIFT (*Scale Invariant Feature Transform*), este seleciona os extremos locais de uma imagem filtrada pelo filtro *Differences of Gaussians* (DoG), uma aproximação do filtro LoG mais rápido de se calcular.

O detector SURF (*Speeded Up Robust Feature*) [Bay et al., 2008] é baseado em um cálculo eficiente do determinante de uma matriz Hessiana. Uma vez que o cálculo da matriz Hessiana implica em convoluções com derivadas Gaussianas de segunda ordem que geram um custo alto, SURF faz uma aproximação com filtros de caixa que podem ser calculados de forma eficiente usando imagens integrais. Aproximando-se, assim, da abordagem com o filtro DoG, porém, com custo computacional reduzido.

Com o objetivo de aproximar mais ainda ao filtro LoG, Agrawal et al. [2008] propuseram o detector CenSurE (*Center-Surround Extrema*). Assim como SURF, CenSurE utiliza filtros de caixa e imagens integrais, porém sua principal diferença é que as características são detectadas em todas as escalas e em todos *pixels* da imagem original. Desta maneira, CenSurE supera as outras abordagens produzindo características mais estáveis [Canclini et al., 2013].

2.1.1.2 Amostragem Densa

Como uma segunda abordagem de seleção dos *patches* a serem descritos, amostragem densa pode ser considerada como a aplicação de uma grade regular sobre a imagem, onde cada célula desta grade será considerada como uma região a ser descrita. A abordagem de amostragem densa oferece uma melhor/maior cobertura de toda cena ou objeto presente na imagem levando a muito mais características locais por imagem. Em

contrapartida, essa abordagem não consegue alcançar o mesmo nível de repetibilidade obtida quando se usa pontos de interesse, a menos que a amostragem seja realizada de uma maneira extremamente densa levando a um alto custo computacional.

Segundo Jurie & Triggs [2005], o uso da abordagem baseada em amostragem densa leva a melhores resultados em tarefas de reconhecimento de objeto e categorização de imagens em geral. Também, foi mostrado por Wang et al. [2009] que a abordagem baseada em amostragem densa supera os resultados por pontos de interesse quando aplicada à detecção de ações em vídeos. No entanto, devido ao custo computacional, alguns trabalhos utilizam uma combinação entre amostragem densa e pontos de interesse [Tuytelaars, 2010; Kim & Grauman, 2011].

2.1.2 Descritores de Características

Uma vez que as regiões a serem descritas de uma imagem foram selecionadas, é necessário descrevê-las de alguma maneira. Idealmente, esta descrição deve ser robusta, concisa e invariante sobre algumas transformações causadas por mudanças de câmera, como variações de rotação, escala, iluminação, entre outras [Ke & Sukthankar, 2004].

Um descritor de características pode ser considerado como uma função aplicada em uma região de uma imagem com o objetivo de descrevê-la. Uma maneira bem simples de se descrever uma região seria representar todos os *pixels*, desta região, em um único vetor. No entanto, dependendo do tamanho da região a ser descrita, isso resultaria em um vetor de alta dimensionalidade levando, também, a uma alta complexidade computacional para um futuro reconhecimento desta região [Mikolajczyk & Schmid, 2005]. Os vetores gerados pelos descritores de características mais comuns na literatura são compostos por valores reais, que são calculados utilizando uma técnica baseada na contagem de ocorrências de orientações de gradiente nas regiões de uma imagem, como: SIFT [Lowe, 2004], HOG (*Histograms of Oriented Gradients*) [Dalal & Triggs, 2005], SURF [Bay et al., 2008].

Nesta seção, é apresentada uma breve revisão dos descritores de características encontrados na literatura, que podem ser classificados de duas maneiras distintas [Canclini et al., 2013]: (i) não-binários e (ii) binários.

É importante ressaltar que novas abordagens para descrição de características vêm aparecendo na literatura, portanto, a lista a seguir não é uma lista exaustiva de todas as abordagens presentes na literatura. No entanto, pode ser considerada como um grupo representativo dos descritores mais relevantes presentes atualmente na literatura.

2.1.2.1 Descritores Não-Binários

SIFT - Scale Invariant Feature Transform

Um dos descritores mais famosos utilizado na literatura é conhecido como SIFT [Lowe, 2004]. Este descritor realiza uma análise espaço-escala levando a um grande desempenho em relação à invariância de escala [Morel & Yu, 2011]. Para cada *patch* a ser descrito, é atribuída uma orientação α selecionando o ângulo que representa o histograma de gradientes locais (calculado para cada *pixel* em torno do *keypoint*). Em seguida, a região de pontos em torno do *keypoint* (centro do *patch* a ser descrito), orientadas por α, é dividida em sub-regiões compostas por um *grid* de tamanho $G \times G$, e um histograma de orientação composto por B *bins* é criado a partir das amostras (suavizadas) de cada sub-região. O descritor é então obtido a partir da concatenação dos histogramas destas sub-regiões, composto pelos $G \times G \times B$ valores. Os valores padrões para G e B geralmente são 4 e 8, respectivamente, resultando em um descritor de 128 dimensões. Por fim, o descritor é normalizado tornando-o robusto a variações de iluminação.

Na literatura, pode-se encontrar várias extensões baseadas no descritor SIFT, como o PCA-SIFT [Ke & Sukthankar, 2004] que aplica *Principal Component Analysis* (PCA) nos pontos/locais a serem descritos, com o objetivo de reduzir o tamanho final do descritor. O descritor GLOH (*Gradient Location and Orientation Histogram*) [Mikolajczyk & Schmid, 2005] foi desenvolvido com base em dois objetivos: aumentar a robustez do descritor SIFT e prover ao descritor um poder maior de distinção. Para isto são extraídos descritores SIFT com um *grid* log-polar e utilizado PCA para a redução da dimensionalidade. Outro exemplo é o descritor HueSIFT [Van de Sande et al., 2010], que apresenta uma concatenação entre um histograma do canal de tonalidade (Hue) ao descritor SIFT, resultando em um descritor com 165 dimensões. Dessa maneira, os autores conseguem adicionar informação de cor ao descritor SIFT. Como outro exemplo, o método chamado RootSIFT, proposto por Arandjelovic & Zisserman [2012], não altera o modo de criação do descritor em si, mas sim a métrica utilizada para calcular as distancias entre eles. De acordo com os autores, o uso da distância de Hellinger traz melhorias ao se fazer o *matching* entre os descritores.

Embora Lowe tenha desenvolvido o descritor SIFT para ser utilizado em aplicações de reconhecimento de objetos, SIFT tornou-se o descritor mais usado em uma infinidade de outras tarefas. Isto se dá devido a seu elevado poder discriminativo e estabilidade.

SURF - Speeded-Up Robust Features

Com o objetivo de superar o problema do alto custo de processamento do SIFT, Bay et al. [2008] propuseram um descritor mais rápido, denominado SURF. Este descritor pode ser visto como uma aproximação do SIFT e possui o mesmo princípio de uso de histogramas baseados em gradientes locais. SURF baseia-se em imagens integrais para aproximar circunvoluções, o que proporciona uma melhoria considerável em termos de eficiência (em comparação com SIFT). Apesar das aproximações na criação do descritor, não existe perda considerável na invariância de rotação e escala.

De maneira similar ao SIFT, o descritor SURF atribui uma orientação para cada *patch* a ser descrito: Uma região circular em torno do *keypoint* é descrita de acordo com a distribuição de respostas obtidas por um filtro Haar-wavelet. O tamanho da região, das wavelets e o parâmetro de amostragem são dependentes de uma escala σ na qual o *keypoint* foi detectado. As respostas de filtro, ponderadas com uma função Gaussiana ao redor do *keypoint*, são representadas por vetores em um espaço bidimensional e então somadas. O maior vetor resultante determina a orientação do *keypoint*. Em seguida, o *patch* é dividido em um *grid* composto por 4×4 sub-regiões. Para cada sub-região, um vetor de características composto por 4 dimensões é calculado usando um filtro Haar-wavelet e então um vetor de soma das orientações é calculado em cada célula. Por fim, a concatenação dos vetores de características, de cada uma das sub-regiões, produz o descritor SURF ($4 \times 4 \times 4 = 64$ dimensões).

2.1.2.2 Descritores Binários

Todos os descritores apresentados anteriormente foram categorizados como descritores não-binários, ou seja, geram vetores de alta dimensionalidade compostos por valores reais. Como valores de ponto flutuante precisam ser codificados por *strings* binárias de 32 *bits* [1], o vetor correspondente a um descritor de 128 dimensões de pontos flutuantes é, de fato, um vetor binário dimensional muito maior. É importante observar também que os descritores SIFT e SURF são baseados em histogramas de gradientes. Desta maneira, os gradientes de cada *pixel* do *patch* precisam ser calculados. Mesmo o descritor SURF acelerando seu cálculo usando imagens integrais, isso ainda não é rápido o suficiente para aplicações que necessitem de um curto tempo de resposta.

Como uma alternativa de baixa complexidade, os descritores binários têm emergido recentemente. Este tipo de descritor tem recebido uma atenção considerável por gerar resultados similares, ou até mesmo melhores, quando comparados a descritores não-binários do estado da arte.

[1] Para a maioria das implementações de variáveis com valores reais.

A ideia básica por trás dos descritores binários é que se pode codificar a maioria das informações de um *patch* em uma sequência binária usando apenas simples testes binários comparando a intensidade entre os *pixels*. Isso pode ser feito de maneira bem rápida, já que apenas comparações de intensidade precisam ser calculadas. Além disso, é possível utilizar a distância de Hamming como medida de distância entre duas *strings* binárias, dessa forma, o *matching* entre dois descritores pode ser feito usando uma única instrução, já que a distância de Hamming é igual à soma da operação XOR entre as duas *strings* binárias.

BRIEF - Binary Robust Independent Elementary Fetures

Como precursor dos descritores binários, BRIEF (*Binary Robust Independent Elementary Features*) [Calonder et al., 2010] pode ser considerado o mais simples dos descritores binários apresentados neste trabalho, pois não possui um padrão de amostragem para descrição de um *patch* nem um mecanismo de compensação de orientação. Basicamente, BRIEF codifica as informações de um *patch* usando apenas simples testes binários, comparando a intensidade entre os *pixels* a partir de um imagem suavizada (por exemplo, usando um núcleo Gaussiano com variância igual a 2 e tamanho igual a 9×9 *pixels*). Por si só, BRIEF não é invariante a escala nem a rotação. No entanto, segundo os autores, seu desempenho é semelhante a um descritor local mais complexo, SURF, quando comparado com a sua robustez à iluminação, borrão, e distorção de perspectiva.

O descritor BRIEF é representado por uma sequência binária construída concatenando os resultados obtidos pelo seguinte teste:

$$b = \begin{cases} 1, & S(p_j) > S(p_i) \\ 0, & \text{caso contrário} \end{cases}$$

onde $S(p_i)$ representa o valor de intensidade do *pixel* no ponto p_i. Apesar das várias formas, apresentadas em [Calonder et al., 2010], para efetuar a seleção de *pixels* que serão comparados, a estratégia mais comum para a escolha destes pontos baseia-se numa forma aleatória de acordo com uma distribuição Gaussiana em volta do *keypoint* de um *patch*. Na Figura 2.2, é ilustrado um exemplo de padrão randômico utilizado na comparação de intensidade entre os *pixels*. Uma observação importante a ser feita é que o número de pontos selecionados implica diretamente no tamanho final do descritor (por exemplo, 128, 256 e 512).

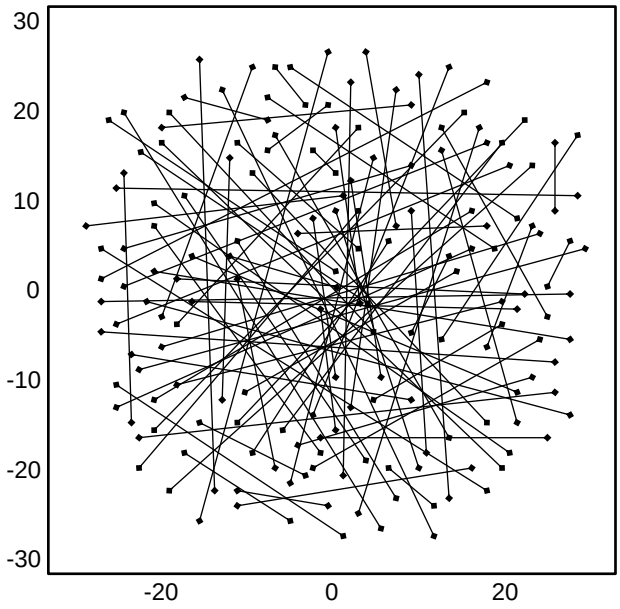


Figura 2.2. Exemplo de padrão randômico utilizado por BRIEF. Cada ponto representa um *pixel* e cada linha representa um simples teste binário comparando a intensidade entre os *pixels*.

ORB - Oriented Fast and Rotated BRIEF

De maneira similar ao BRIEF, o descritor ORB (*Oriented Fast and Rotated BRIEF*) [Rublee et al., 2011] pode ser considerado como uma alternativa ao SIFT e SURF. O descritor ORB é robusto a ruídos e invariante à rotação, resolvendo assim um dos problemas de seu antecessor BRIEF. Apesar desta melhoria, o descritor ORB permanece apenas parcialmente invariante à escala.

De acordo com os autores, a invariância à rotação é obtida estimando a rotação do *patch* usando a intensidade do centroide, calculado a partir de *patch moments* [Rosin, 1999] no qual supera abordagens baseadas em gradientes. Em seguida, é aplicada a estimação de orientação ao *patch* a ser descrito e o descritor é construído da mesma maneira que BRIEF, usando os simples testes binários de comparação de intensidade entre os *pixels*. Finalmente, um subconjunto de testes binários é escolhido de modo a reduzir a sua inter-relação, aumentando assim o poder de discriminação do descritor. O algoritmo funciona com uma abordagem gulosa, selecionando os pares de amostragem com a maior variância, parando quando 256 testes binários são selecionados, implicando na dimensionalidade final do descritor.

BRISK - Binary Robust Invariant Scalable Keypoints

Assim como BRIEF e ORB, o descritor BRISK (*Binary Robust Invariant Scalable Keypoints* [Leutenegger et al., 2011] encaixa-se na categoria de descritores binários. A abordagem utilizada por BRISK é muito semelhante ao BRIEF no sentido em que o descritor é calculado com base em simples testes binários de comparação de intensidade entre os *pixels*. Porém, BRISK possui três principais diferenças quando comparado ao BRIEF: (i) leva em consideração a rotação de um ponto a ser descrito; (ii) faz uso de uma teoria escala-espaço para adaptar o padrão de amostragem ao máximo no espaço de escala; e (iii) usa um padrão especial para os testes binários, em vez de uma distribuição aleatória. Desta maneira, BRISK se torna invariante à rotação e escala.

Como ilustrado na Figura 2.3, o descritor BRISK usa um padrão de pontos p_i igualmente distribuídos em círculos concêntricos ao redor do *keypoint* do *patch* a ser descrito. Para comparação dos pontos, os autores definem dois conjuntos distintos de pares de pontos, pares de longa distância e pares de curta distância. Os pares de longa distância são compostos por pares (i, j), em que $\|p_i - p_j\| > \delta_{min}$, e são utilizados para estimar a orientação do *keypoint* usando uma média de gradiente. Em seguida, é aplicada uma suavização Gaussiana sobre os anéis concêntricos utilizados pelo padrão de amostragem e 512 pares de curta distância (cuja distância é menor que um limiar δ_{max}) são utilizados para construção do descritor.

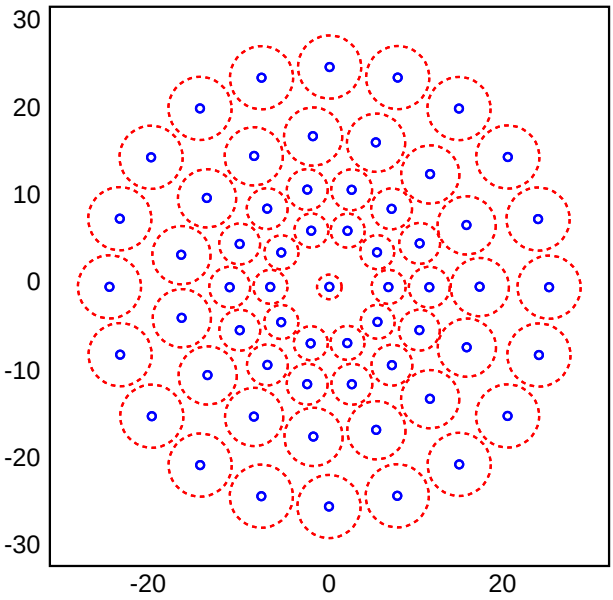


Figura 2.3. Padrão de amostragem do descritor BRISK baseado em 60 pontos: os pequenos círculos azuis indicam os locais de amostragem; os círculos maiores tracejados em vermelho correspondem a um desvio padrão com núcleo Gaussiano utilizado para suavizar a intensidade dos valores nos pontos de amostragem.

FREAK - Fast Retina Keypoint

Com o objetivo de estender o descritor BRISK, Ortiz [2012] propôs um descritor denominado FREAK (*Fast Retina Keypoint*). Como no caso do descritor BRISK, FREAK tem seu padrão de amostragem baseado em Gaussianas, porém a distribuição de amostragem dos pontos é biologicamente inspirada no padrão da retina do olho humano.

De acordo com Ortiz, FREAK apresenta duas diferenças importantes em relação a BRISK. A primeira consiste em uma alocação de distribuições concêntricas com um crescimento exponencial em relação à distância do *keypoint*. A segunda contribuição é baseada no fato do padrão de amostragem, criando sobreposições sobre diferentes círculos concêntricos, como apresentado na Figura 2.4. A sobreposição entre as regiões de amostragem acrescenta redundância que aumenta o poder discriminativo do descritor. Segundo o autor, essa redundância também está presente nos campos receptivos da retina humana.

O padrão de amostragem utilizado na implementação original do descritor FREAK é composto por 43 "campos receptivos", levando a 903 testes de comparações possíveis. Portanto, para a construção do descritor final, FREAK utiliza uma abordagem similar ao descritor ORB, selecionando com uma abordagem gulosa os testes de comparação menos correlacionados e, portanto, mais discriminativos. Para obter máximo desempenho, são utilizados 512 testes binários.

BinBoost

Recentemente, Trzcinski et al. [2013] propuseram um novo *framework* com o objetivo de gerar um descritor binário extremamente compacto e altamente discriminativo. Denominado BinBoost, este descritor é robusto a mudanças de iluminação e ponto de vista.

Diferente dos descritores binários mencionados anteriormente, que calculam o descritor final com base em simples testes binários comparando a intensidade entre *pixels*, cada *bit* gerado pelo BinBoost é calculado usando uma função Hash binária da mesma forma que o classificador AdaBoost [Freund & Schapire, 1997]. Essa função é baseada em *weak learners* que levam em consideração orientações de gradientes de intensidade sobre o *patch* a ser descrito. A função Hash é otimizada de forma iterativa, ou seja, a cada iteração, amostras incorretas serão atribuídas a um peso maior, enquanto o peso das amostras corretas será diminuído. Desta maneira, o próximo *bit* a ser calculado tenderá a corrigir o erro de seus antecessores.

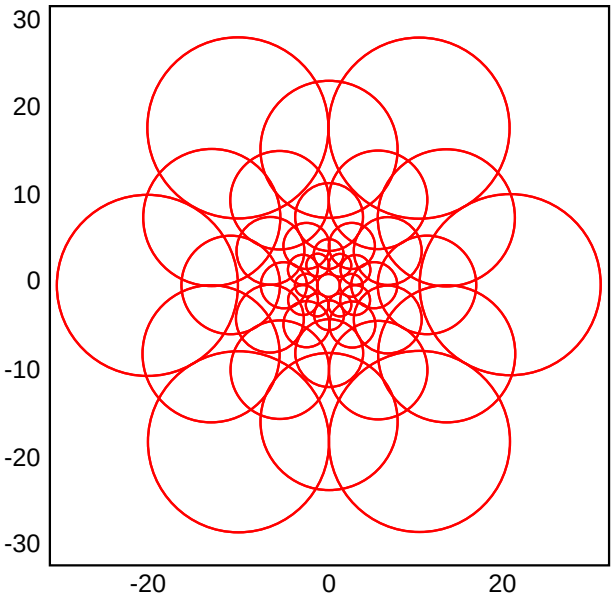


Figura 2.4. Padrão de amostragem usado pelo descritor FREAK. Cada círculo vermelho representa um "campo receptivo" onde um desvio padrão com núcleo Gaussiano é utilizado para suavizar a intensidade dos valores nos pontos de amostragem.

2.2 Representações Intermediárias

Seguindo a abordagem mencionada no início deste capítulo, com as características locais já extraídas em mãos, torna-se necessário codificá-las de alguma maneira para que se tenha uma representação global da imagem ou vídeo baseada em características locais. Essa codificação é denominada na literatura como representação intermediária ou representação *mid-level*.

A abordagem de representação intermediária mais utilizada na literatura, inicialmente proposta por Sivic & Zisserman [2003], é conhecida como *Bag-of-Words* (BoW) ou *Bag-of-Features* (BoF). A ideia básica por trás do modelo BoW consiste em quantizar as características locais obtidas em palavras visuais, de acordo com um dicionário visual pré-construído, e então representar cada imagem como um vetor composto por estas palavras visuais.

Várias extensões do modelo BoW foram propostas na literatura. Entre elas, ressalta-se a recente representação intermediária proposta por Avila et al. [2013]. Denominada BossaNova, esta enriquece o modelo BoW mantendo um histograma de distâncias entre os descritores encontrados na imagem, em relação a cada palavra visual presente no dicionário visual.

A seguir, as duas representações intermediárias mencionadas anteriormente são

detalhadas. Mais informações sobre outras representações intermediárias podem ser encontradas em [Lazebnik et al., 2006; Boureau et al., 2010; Chatfield et al., 2011].

2.2.1 Bag-of-Words

Segundo Boureau et al. [2010], o modelo BoW pode ser compreendido como a aplicação de duas etapas críticas : codificação e *pooling*. A etapa de codificação quantifica as características locais extraídas da imagem de acordo com um dicionário visual, também conhecido como *codebook*, associando os descritores locais extraídos da imagem com o elemento mais próximo deste vocabulário visual. O dicionário visual normalmente é construído aplicando um algoritmo de clusterização, geralmente *k-means* [Lloyd, 2006], em um conjunto de amostras dos descritores locais extraídos, onde cada palavra visual (*codewords*) corresponde ao centroide obtido de cada *cluster*. A etapa de *pooling* resume as palavras visuais obtidas em um único vetor de características com o objetivo de representar toda a imagem.

A seguir, é apresentada a formulação matemática utilizada por Avila [2013] para descrever o modelo BoW. Considere $\mathcal{X} = \{\mathbf{x}_j\}$, $j \in \{1, \ldots, N\}$ como um conjunto não ordenado de características locais de uma imagem, onde $\mathbf{x}_j \in \mathbb{R}^D$ é um vetor gerado por um descritor local e N é o número de características locais extraídas de uma imagem. Considere $\mathcal{C} = \{\mathbf{c}_m\}$, $m \in \{1, \ldots, M\}$ como um dicionário visual, onde $\mathbf{c}_m \in \mathbb{R}^D$ corresponde a uma palavra visual e M corresponde ao número total de palavras visuais.

A etapa de codificação pode ser entendida como uma função de ativação f para o dicionário visual, ativando apenas a palavra visual mais próxima a característica local, atribuindo peso zero a todas as outras:

$$\begin{aligned} f : \mathbb{R}^D &\longrightarrow \mathbb{R}^M, \\ \mathbf{x}_j &\longrightarrow f(\mathbf{x}_j) = \alpha_{\mathbf{j}} = \{\alpha_{m,j}\}, \qquad m \in \{1, \ldots, M\}, \\ \alpha_{m,j} &= \begin{cases} 1, & \text{sse} \quad m = \arg\min \|x_j - c_k\|_2^2, \quad k \in \{1, \ldots, M\} \\ 0, & \text{caso contrário} \end{cases} \end{aligned} \tag{2.1}$$

onde $\alpha_{m,j}$ é a componente m do vetor codificado α_j. Esse esquema de codificação é mencionado na literatura como *hard coding* sobre o dicionário visual.

Em seguida, a etapa de *pooling* pode ser representada pela função g:

$$
\begin{aligned}
g : \mathbb{R}^N &\longrightarrow \mathbb{R}, \\
\alpha_j = \{\alpha_{m,j}\}, j \in \{1, \ldots, N\} &\longrightarrow g(\{\alpha_j\}) = \mathbf{z} : \forall m, z_m = \sum_{j=1}^{N} \alpha_{m,j}
\end{aligned}
\tag{2.2}
$$

Por fim, a representação final da imagem é dada pelo vetor: $\mathbf{z} = [z_1, z_2, \ldots, z_M]^T$. Na Figura 2.5, é ilustrado todo processo de classificação usado pelo modelo BoW.

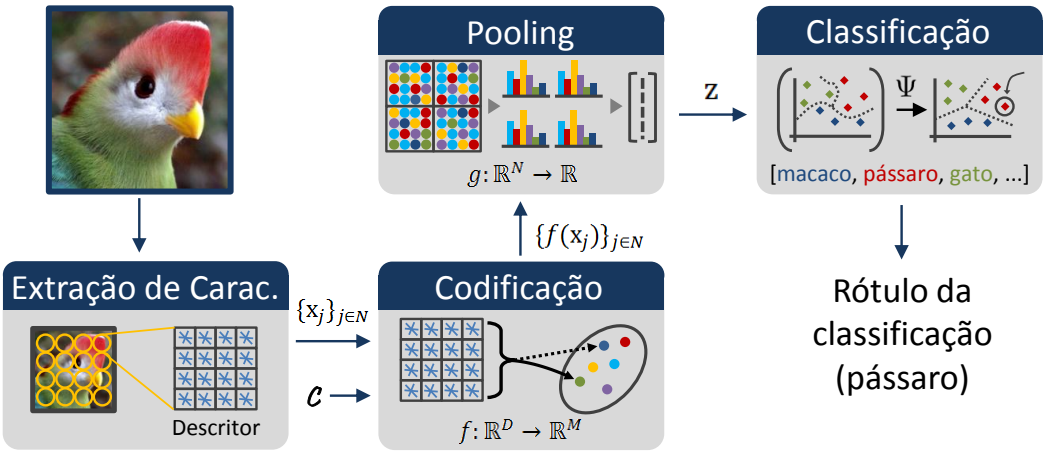


Figura 2.5. Processo de classificação usado pelo modelo *Bag-of-Words*. Primeiro os descritores locais são extraídos da imagem, $\{\mathbf{x}_j\}_{j \in N}$, onde $\mathbf{x}_j \in \mathbb{R}^D$. Depois, na fase de codificação, uma função f ativa a palavra visual mais próxima ao descritor local, atribuindo peso zero a todas as outras. Em seguida, na etapa de *pooling*, a função g resume as palavras visuais obtidas em um único vetor de características $\mathbf{z}$. Por fim, um algoritmo de classificação (por exemplo, SVM) é treinado com base nos vetores BoW obtidos. Imagem adaptada de Chatfield et al..

2.2.2 BossaNova

Como uma extensão do modelo BoW, a representação intermediária BossaNova [Avila et al., 2013] oferece um aprimoramento na etapa de *pooling*, a fim de preservar de uma maneira mais rica a informação obtida durante a etapa de codificação. Desta maneira, em vez de compactar toda a informação relacionada a uma palavra visual em um único valor escalar, a etapa de *pooling* resulta em uma distribuição de distâncias. Para isto, Avila et al. usaram uma estimação não-paramétrica da distribuição dos descritores,

calculando um histograma de distâncias entre os descritores encontrados na imagem e cada palavra visual presente no dicionário visual.

Neste trabalho, é apresentada apenas uma breve introdução da representação intermediária BossaNova. Mais detalhes podem ser encontrados em [Avila et al., 2011, 2013].

Considere $\mathcal{X} = \{\mathbf{x}_j\}$, $j \in \{1, \ldots, N\}$ como um conjunto não ordenado de características locais de uma imagem, onde $\mathbf{x}_j \in \mathbb{R}^D$ é um vetor gerado por um descritor local e N é o número de características locais extraídas de uma imagem. Considere $\mathcal{C} = \{\mathbf{c}_m\}$, $m \in \{1, \ldots, M\}$ como um dicionário visual, onde $\mathbf{c}_m \in \mathbb{R}^D$ corresponde a uma palavra visual e M corresponde ao número total de palavras visuais.

A representação intermediária BossaNova segue o formalismo BoW (codificação/*pooling*), mas propõe uma representação da imagem que mantém mais informações do que BoW durante a etapa de *pooling*. Assim, na etapa de codificação do BossaNova, Avila et al. utilizam uma estratégia *soft coding*, considerando as "K-palavras visuais" mais próximas para codificação de um descritor local. Matematicamente falando, a etapa de codificação de BossaNova pode ser modelada por uma função f da seguinte forma:

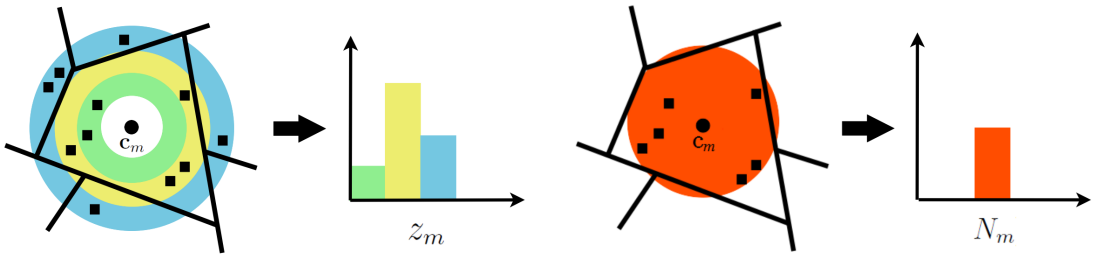
$$\begin{aligned} f : \mathbb{R}^D &\longrightarrow \mathbb{R}^M, \\ \mathbf{x}_j &\longrightarrow f(\mathbf{x}_j) = \alpha_{\mathbf{j}} = \{\alpha_{m,j}\}, \\ \alpha_{m,j} &= \frac{exp^{-\beta_m d_2(\mathbf{x}_j, \mathbf{c}_m)}}{\sum_{m'=1}^{K} exp^{-\beta_m d_2(\mathbf{x}_j, \mathbf{c}_{m'})}} \end{aligned} \tag{2.3}$$

onde $d_2(\mathbf{x}_j, \mathbf{c}_m)$ é a distância entre $\mathbf{c}_m$ e $\mathbf{x}_j$. O parâmetro β_m é um regulador para o *soft coding* (quanto maior, mais próximo de *hard coding*).

A etapa de *pooling* pode ser modelada pela função g, que estima a função de densidade de probabilidade de α_m: $g(\alpha_m) = \text{fdp}(\alpha_{\mathbf{m}})$, calculando o seguinte histograma de distâncias $z_{m,b}$:

$$
\begin{aligned}
g : \mathbb{R}^N &\longrightarrow \mathbb{R}^B, \\
\alpha_{\mathbf{m}} &\longrightarrow g(\alpha_m) = z_m, \\
z_{m,b} &= \operatorname{card}\left(\mathbf{x}_j \mid \alpha_{m,j} \in \left[\frac{b}{B}; \frac{b+1}{B}\right]\right), \\
&\quad \frac{b}{B} \geq \alpha_m^{min} \quad \text{and} \quad \frac{b+1}{B} \leq \alpha_m^{max},
\end{aligned}
\tag{2.4}
$$

onde B representa o número de *bins* de cada histograma z_m, e $[\alpha_m^{min}; \alpha_m^{max}]$ limita o intervalo de distâncias para os descritores considerados no cálculo do histograma. Uma comparação entre as funções de *pooling* do modelo BoW e BossaNova é ilustrada na Figura 2.6.



(a) Função *pooling* BossaNova (b) Função *pooling Bag-of-Words*

Figura 2.6. Ilustração das funções *pooling* do modelo *Bag-of-Words* e BossaNova. (a) A função *pooling* de BossaNova representa a distribuição de densidade das distâncias ($B = 3$ *bins*) entre a palavra visual c_m e os descritores locais da imagem. Para cada centro c_m é obtido um histograma local z_m. Cada anel colorido representa uma distância entre o centro c_m e os descritores locais (representados por pontos pretos). Para cada *bin* colorido $z_{m,b}$, seu valor é igual ao número de descritores locais cuja distância da palavra visual c_m cai no *bin* b. (b) Caso $B = 1$, o histograma z_m se reduz a um único escalar N_m contando o número de características locais próximas ao centro c_m. Imagem de Avila et al. [2012].

Após o cálculo dos histogramas z_m para todos os centros $\mathbf{c}_m$, o vetor BossaNova $\mathbf{z}$ pode ser escrito da seguinte maneira:

$$
\mathbf{z} = \left[\left[z_{m,b}\right], st_m\right]^{\mathrm{T}}, \quad (m,b) \in \{1,\ldots,M\} \times \{1,\ldots,B\}, \tag{2.5}
$$

onde $\mathbf{z}$ é um vetor de tamanho $M \times (B+1)$, s é uma constante não negativa e t_m é um valor escalar para cada palavra visual, contando o número de descritores $\mathbf{x}_j$ próximos àquela palavra visual.

Em resumo, a representação BossaNova é definida por três parâmetros: M, o número de palavras visuais; B, o número de *bins* para cada histograma; e $[\alpha_m^{min}, \alpha_m^{max}]$, o intervalo de distâncias – distância mínima, α_m^{min}; e distância máxima, α_m^{max}; no espaço de descritores $\mathbb{R}^D$ que definem os limites do histograma.

Avila et al. aplicaram a representação BossaNova no contexto de reconhecimento de objetos e detecção de pornografia. Em comparação ao modelo BoW, BossaNova se sobressai de maneira significante [Avila et al., 2011, 2012, 2013], apenas usando um simples histograma de distâncias para capturar as informações relevantes. BossaNova mostra ser um método muito flexível, mantendo uma representação final bem compacta.

2.3 Classificação - Aprendizado de Máquina

Segundo Ghahramani [2004], aprendizado de máquina é o campo de pesquisa dedicado ao estudo formal de sistemas de aprendizagem. Pode ser considerado como um campo altamente interdisciplinar por se basear em ideias de diversas áreas, como estatística, ciência da computação, engenharia, ciência cognitiva, teoria de otimização, entre outras. Técnicas de aprendizado de máquina podem ser separadas em várias categorias (supervisionado, não-supervisionado, semi-supervisionado, ativo, meta aprendizado), porém, de uma forma geral, a distinção mais fundamental é entre algoritmos de aprendizado supervisionado e não-supervisionado.

No caso de aprendizado supervisionado, a "máquina" recebe uma sequência de rótulos como saídas desejadas e o objetivo desta "máquina" é aprender a produzir uma saída correta dada uma nova entrada. No caso de uma tarefa de classificação, esta saída pode ser um rótulo pertencente à classe predita.

Um outro tipo principal de algoritmos de aprendizado de máquina é o aprendizado não-supervisionado. Neste caso, a "máquina" simplesmente recebe as entradas, sem obter nenhuma sequência de rótulos como saídas desejadas, ou seja, nenhum *feedback* do ambiente. De certo modo, o aprendizado não-supervisionado pode ser visto como um algoritmo cujo objetivo é encontrar padrões nos dados de entrada para poder então separá-los [Ghahramani, 2004]. Algoritmos de clusterização podem ser vistos como um exemplo clássico de aprendizado não-supervisionado.

A seguir, é apresentada uma breve introdução do algoritmo de aprendizado supervisionado SVM. Esta técnica tem sido amplamente utilizada pela literatura em diversas tarefas que envolvam reconhecimento visual, como por exemplo classificação de imagens [Lazebnik et al., 2006; Gemert et al., 2008; Yang et al., 2009; Perronnin et al., 2010; Zhou et al., 2010; Krapac et al., 2011; SáNchez et al., 2012; Avila et al.,

2013], reconhecimento de ações [Schuldt et al., 2004; Laptev et al., 2008; Kläser et al., 2008; Wang et al., 2009; Gaidon et al., 2012; Jhuang et al., 2013] e detecção de pornografia [Deselaers et al., 2008; Lopes et al., 2009b,a; Ulges & Stahl, 2011; Avila et al., 2011, 2013; Caetano et al., 2014].

Mais detalhes sobre técnicas de aprendizado de máquina podem ser encontradas em [Ghahramani, 2004; Özgür, 2004; Zhu, 2006; Kotsiantis, 2007].

2.3.1 Support Vector Machines

A técnica *Support Vector Machines* (SVM) [Vapnik, 1995, 1998], é utilizada em muitas tarefas que envolvem aprendizado de máquina, como reconhecimento de padrões e reconhecimento visual, devido à sua alta capacidade de generalização e robustez contra ruídos e *outliers*. Primeiramente, o SVM foi desenvolvido como uma máquina de decisão binária, ou seja, suportando apenas duas classes [Vapnik & Lerner, 1963]. O método foi proposto pela primeira vez como um classificador linear, mas foi então estendido para lidar com problemas não-linearmente separáveis usando funções de *kernel* [Aizerman et al., 1964].

Segundo Bkassiny et al. [2013], a ideia básica do SVM consiste em mapear os vetores de entrada para um espaço de características de alta dimensionalidade em que esses vetores de entrada se tornem linearmente separáveis. Este mapeamento, do espaço vetorial de entrada para o espaço de características, é um mapeamento não-linear que pode ser feito por meio de funções de *kernel*. Alguns *kernels* populares são: Linear, Polinomial e Gaussiano. A habilidade de separar dados com distribuição não-linear está relacionada com a escolha dessa função, e que deve ser analisada de acordo com o domínio do problema [Duda & Hart, 2000].

Durante a classificação, o objetivo é encontrar um hiperplano que permita uma maior generalização no espaço de alta dimensionalidade. Este hiperplano é chamado classificador de margem máxima. Pode ser que existam diferentes hiperplanos possíveis que separem as duas classes de dados, porém, apenas um deles permitirá uma margem máxima. A margem é a distância a partir de um hiperplano de separação em relação aos dados. Esses dados mais próximos são chamados de vetores de suporte e o hiperplano que permite a margem máxima é chamado de hiperplano de separação ótimo. Uma ilustração é apresentada na Figura 2.7.

Uma introdução mais profunda e abrangente sobre SVMs pode ser encontrada em [Cristianini & Shawe-Taylor, 2000; Scholkopf & Smola, 2001].

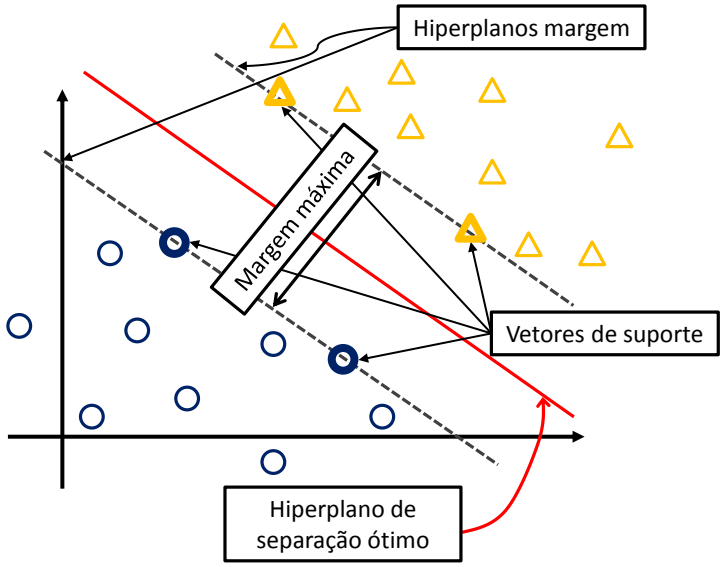


Figura 2.7. Ilustração da ideia básica de classificação binária com SVM: hiperplano de separação ótimo (linha vermelha sólida) e dois hiperplanos margem (linhas tracejadas); os vetores de suporte são os dados em negrito.

Capítulo 3

Revisão da Literatura

Segundo Ries & Lienhart [2014], os trabalhos da literatura que envolvem detecção de pornografia podem ser divididos em três grupos principais: (i) abordagens baseadas em cor de pele, que exploram a hipótese de que imagens/vídeos pornográficos geralmente apresentam grandes áreas com cores de pele; (ii) abordagens baseadas em informação de forma; e (iii) abordagens baseadas em características locais em conjunto com modelos *Bag-of-Words*.

A seguir, é apresentada uma revisão da literatura de trabalhos que envolvem a detecção de pornografia. Na Seção 3.1, são abordados os trabalhos baseados em detecção de pele, assim como os trabalhos baseados em informação de forma, um vez que todas as abordagens baseadas em forma, apresentadas neste trabalho, também contam com a etapa de encontrar *pixels* que apresentem cores relacionadas à pele. Na Seção 3.2, são apresentados os trabalhos que fazem uso de características locais e modelos *Bag-of-Words* (BoW). Será apresentada, ainda, a Seção 3.3 que agrupa trabalhos que abordam diferentes técnicas, como por exemplo informação de movimento e análise de áudio.

3.1 Abordagens Baseadas em Cor de Pele e Forma

A maioria dos trabalhos de detecção de pornografia são feitos baseados em abordagens de detecção de cor de pele e forma. Isto se dá pelo fato de que a propriedade mais óbvia em imagens pornográficas é uma grande fração de *pixels* que apresentam cores relacionadas à pele. Também, a maioria das imagens pornográficas compartilha de algumas formas características [Ries & Lienhart, 2014].

As abordagens apresentadas em [Forsyth & Fleck, 1996, 1997, 1999] começam encontrando regiões com cores de pele na imagem. Para isto, transformam cada valor do

pixel em um valor de intensidade e dois valores de tonalidade. Feito isso, são aplicadas regras de decisão a fim de encontrar regiões com cores da pele. Após a detecção da região de pele, é aplicado um detector de canto e uma transformada de Hough para encontrar candidatos a membros humanos. Estes candidatos são iterativamente combinados de acordo com um conjunto de restrições que modela a geometria do corpo humano. Se for possível reunir os membros de uma forma geometricamente razoável, a imagem é classificada como pornográfica.

Jones & Rehg [2002] construíram um histograma 3D com 256 *bins* para cada canal de cor. A partir destes histogramas, são extraídas cinco características diferentes para cada imagem, como por exemplo, a porcentagem de *pixels* relacionados à pele ou o número de áreas de pele conectadas. Por fim, é então treinada uma árvore de decisão baseada nessas características. No entanto, os autores mostram resultados sugerindo que histogramas com 32 *bins*, para cada cor, são suficientes e superam até mesmo histogramas mais detalhados.

Em [Zheng et al., 2004], os autores estimam a probabilidade dos *pixels* da imagem serem relacionados à pele usando um modelo de entropia máxima. Eles determinam a distribuição de probabilidade com a entropia máxima com repeito a restrições vindas de uma fase de treinamento. Uma vez que cada cor possível representa uma restrição para o modelo de entropia máxima, o número de parâmetros é enorme. Portanto, os parâmetros são estimados utilizando árvores Bethe[1]. Como saída, obtém-se um "mapa de peles" em escalas de cinza, com as escalas de cinza sendo proporcionais às probabilidades de pele. A partir deste "mapa de peles", são extraídas características, como em [Jones & Rehg, 2002], e por fim uma rede neural é utilizada como classificador final.

Inspirados nos histogramas de cor de Jones & Rehg, Rowley et al. [2006] geram um mapa baseado em cor de pele para a imagem e então determinam componentes conectados neste mapa. Em seguida, são extraídas características do mapa baseado em pele e também dos componentes conectados, como média e desvio padrão. Além disso, os autores também utilizam outras características de cor, como os *pixels* de borda dentro das regiões de pele. Por fim, essas características são utilizadas como entrada para um classificador *Support Vector Machines* (SVM).

Lee et al. [2007] apresentam uma abordagem que utiliza um esquema de aprendizado baseado na distribuição cromática de pele na imagem, utilizando uma rede neural para aprender e julgar se a imagem de entrada contém exposição de pele e então segmentá-las. Além disso, é utilizada uma característica para detectar texturas

[1] Árvores Bethe são estruturas de grafo capazes de simular a vizinhança entre os *pixels* de uma forma livre de parâmetros.

com rugosidade a fim de rejeitar objetos que não contenham pele. Em seguida, são extraídas três tipos de características relacionadas à forma segmentada (tamanho da área ocupada, razão de aspecto e localização) e enviadas a um classificador AdaBoost [Freund & Schapire, 1997]. Por fim, é aplicado um algoritmo de detecção de faces para filtrar falsos candidatos relacionados a fotos de rosto (como por exemplo, fotos 3x4 que apresentam uma grande quantidade de pele).

Em [Hu et al., 2007], é apresentado um *framework* para detectar páginas da Web com conteúdo pornográfico, baseado em características de texto e imagens. Para isto, as imagens são divididas em blocos retangulares e então é analisado se dentro de cada bloco existe uma quantidade de *pixels*, relacionados à pele, acima de um limiar. Caso isso seja verdadeiro, é retornada uma região conectada por esses blocos. Os cantos interiores dos blocos da maior região conectada são então considerados como pontos de interesse. Por fim, são extraídas características destes pontos de interesse, como quantidade de *pixels* de pele e características relacionadas à forma, e então enviadas a um classificador.

Em [Wu et al., 2008], os autores adicionaram ao *framework* de Hu et al. um filtro baseado em cor com o objetivo detectar a presença humana em imagens. Em [Zuo et al., 2010], os mesmos autores introduziram um detector de regiões baseado em *pixels* de pele, combinando 31 tipos de características diferentes. Já em [Hu et al., 2011], o *framework* foi alterado para usar uma abordagem de detecção de regiões com padrões de pele, em vez de *pixels* de pele, resultando em melhores resultados.

Lee et al. [2013] propõem um sistema de detecção de imagens pornográficas composto por três etapas. A primeira etapa é usada para separar as cores da imagem em grupos de pele e não-pele, para isto é utilizado o esquema proposto em [Lee et al., 2007]. Na segunda etapa, é feita uma análise de textura para verificar a probabilidade da região ser composta por pele ou não. Na ultima etapa, é aplicado um algoritmo de detecção de faces para eliminar fotos de rosto. Além disso, Lee et al. verificaram a presença de "buracos" nas imagens binarizadas para detectar fotos relacionadas às roupas de banho. Para as imagens restantes, são extraídas características relacionadas à posição da região de pele e características morfológicas e então usadas para treinar um classificador SVM.

Um novo método para estimar regiões de pele foi proposto por Yu & Han [2014] usando simples operações no espaço de cor HSV (*Hue, Saturation, Value*) mais um pós-processamento adicional para reduzir ruídos. O método mostra-se rápido e com bons resultados suficientes para ser usado como filtragem de imagens pornográficas claras, antes de um processo de identificação mais robusto. Basicamente, é utilizado um limiar para selecionar *pixels* relacionados a pele no componente *Hue*. Em seguida,

para remover regiões detectadas incorretamente, é calculado um mapa de densidade de bordas para a imagem. Os autores então usam a premissa de que a densidade de bordas é baixa em regiões de pele, desta maneira, são removidos os *pixels* que apresentam alta densidade de borda. Além disso, são utilizadas operações morfológicas com o objetivo de reduzir possíveis ruídos. Por fim, média e desvio padrão das regiões de pele são calculados e então é utilizado um outro limiar para decidir se a imagem é pornográfica ou não.

Apesar de existirem muitas abordagens baseadas em detecção de cor de pele para classificar conteúdos pornográficos, essas abordagens geralmente têm como desvantagem uma alta taxa de falsos positivos, pois nem todas imagens com grandes áreas de exposição de pele são necessariamente pornográficas (imagens com pessoas usando roupas de banho, ou imagens relacionadas a esportes). Além disso, um outro obstáculo é a diversidade de cor de pele humana existente, dificultando ainda mais o processo de classificação. Outro problema a ser observado é que imagens em escalas de cinza não podem ser classificadas usando características relacionadas à cor [Ries & Lienhart, 2014]. Abordagens relacionadas às características de forma apresentam o mesmo problema, pois também utilizam informações de cor de pele.

3.2 Abordagens Baseadas em Características Locais

Outra abordagem utilizada na literatura para detecção de pornografia são trabalhos que empregam a extração de características locais da imagem. A maioria destes trabalhos utilizam o modelo BoW, ou alguma de suas extensões, como representação intermediária para codificar as características locais.

Deselaers et al. [2008] foram os primeiros a utilizar características locais em conjunto com modelos BoW. Os autores propuseram uma abordagem baseada no modelo BoW para filtrar e classificar pornografia em diferentes categorias. Para a detecção das características locais, foi utilizado o detector SIFT. Com os pontos detectados, cada *pacth* é reduzido utilizando PCA. Deselaers et al. não utiliza nenhum descritor para descrever as regiões detectadas, alegando vantagem dos *patches* por fornecerem informação de cor. Para a etapa de classificação, é utilizado um classificador SVM.

Seguindo a ideia anterior, Lopes et al. apresentam uma abordagem utilizando o modelo BoW em conjunto com descritores HueSIFT. Em Lopes et al. [2009b], os autores realizam a classificação de imagens utilizando o detector SIFT, descritores HueSIFT e um classificador SVM. No mesmo trabalho, também é feita uma comparação entre

os descritores SIFT e HueSIFT aplicados à pornografia, mostrando que a combinação entre descritor e informação de cor produz melhores resultados. Em Lopes et al. [2009a], os autores estendem seu trabalho para detecção em vídeos. Para isto, realizam a mesma abordagem anterior para quadros selecionados do vídeo. Por fim, é realizada uma votação majoritária em cima da classificação final, dos quadros, para definir a classe final do vídeo.

Em contraste com as abordagens anteriores, Ulges & Stahl [2011] realizam experimentos em colaboração com a polícia para detectar pedofilia. Como características locais, foram utilizados os coeficientes de baixa frequência da transformada discreta do cosseno em *patches* de imagens no modelo de cor YUV. Para codificação das características locais, também foi utilizado o modelo BoW, e, como nas abordagens anteriores, SVM para classificação. Os autores também implementaram uma abordagem baseada em cor, similar à de Jones & Rehg [2002] com o objetivo de comparação, chegando à conclusão que a abordagem baseada em BoW supera abordagens baseadas apenas em cor.

Steel [2012] propôs um método de reconhecimento de imagens pornográficas baseado em palavras visuais. Para isto, foi proposto uma variação do descritor SIFT, Mask-SIFT, que usa um pré-filtro Gaussiano para remover todos os *pixels* de uma imagem que não são relacionados à pele. A imagem é então processada usando um filtro da mediana para preencher *pixels* em falta e eliminar ruídos, criando uma "imagem máscara". Uma vez que a "imagem máscara" é criada, o descritor SIFT é usado para extrair características a partir das partes relacionadas às pessoas da imagem. Em seguida, as características são agrupadas em palavras visuais. Por fim, Steel utiliza um classificador baseado em cascata que filtra as imagens baseado em tom de pele, forma e características locais para determinar se uma imagem é pornográfica.

Em [Avila et al., 2013], os autores apresentam uma abordagem para classificação de vídeos pornográficos. Primeiramente, os vídeos são segmentados em tomadas e então é feita uma extração do quadro central de cada tomada para representar o vídeo. Em seguida, são extraídas características locais com o descritor HueSIFT de maneira densa (amostragem densa). Para a representação dos descritores, é utilizado o meio de representação intermediária BossaNova. Um classificador SVM é então utilizado para classificar os quadros centrais extraídos de cada tomada. Por fim, uma votação majoritária é utilizada para predizer a classe do vídeo. A Figura 3.1 ilustra a abordagem apresentada por Avila et al..

Segundo Ries & Lienhart [2014], de uma forma geral, abordagens baseadas em características locais têm mostrado resultados mais satisfatórios do que as abordagens baseadas em informações de cor. Uma importante vantagem do uso de abordagens com

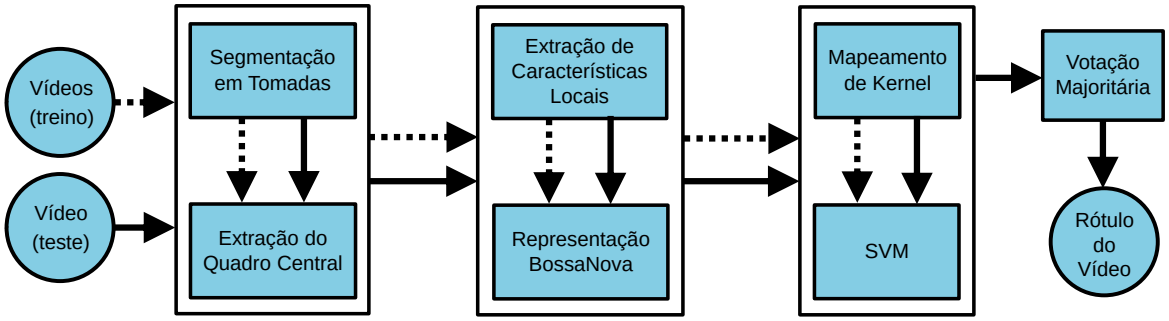


Figura 3.1. Abordagem para a classificação de vídeos pornográficos [Avila et al., 2013]. A fase de treinamento é representada pelas linhas tracejadas, enquanto a etapa para a classificação é representada pelas linhas contínuas. Imagem adaptada de [Avila, 2013].

características locais é que essas características podem ser calculadas independentes da informação de cor. Outra vantagem é que elas modelam, de forma compacta, as regiões da imagem em um vetor de tamanho fixo, facilitando a comparação de regiões de imagem, assim como da imagem como um todo. No entanto, a extração de características locais pode vir a ser mais demorada do que examinar características de imagem relacionadas à cor.

3.3 Outras Abordagens

Além das abordagens de detecção de pornografia apresentadas anteriormente, podem ser encontrados, na literatura, trabalhos que exploram diferentes abordagens, como informação de movimento (características espaço-temporais) e abordagens que envolvem análise de áudio.

Um método baseado em vetores de movimento para classificação de conteúdos de vídeo foi proposto por Endeshaw et al. [2008]. Primeiramente, o vídeo é dividido em pequenos segmentos de tamanho fixo e então é criado um vetor de movimento dominante para cada quadro. O algoritmo então tenta detectar movimentos repetitivos em uma faixa de frequência específica durante 16 segundos de intervalo usando estimação espectral. Em uma segunda etapa, um limiar é então utilizado para determinar se movimentos repetitivos durante um longo período podem ser classificados como material pornográfico.

Jansohn et al. [2009] apresentam um *framework* para detectar conteúdos pornográficos em vídeos. Jansohn et al. calculam repetições de movimentos, chamados detecção de periodicidade (PER), naturalmente envolvidos em vídeos pornográficos, gerando histogramas de movimento. Os autores combinam análises de quadros, in-

cluindo modelos BoW e detecção de pele, com os resultados de análise de movimento. A combinação é feita usando uma fusão tardia com os *scores* dos classificadores usados em cada análise.

Em Valle et al. [2012], é realizada uma comparação entre características de cor, com histogramas no modelo RGB (*Red, Green, Blue*), características locais (descritores SIFT e HueSIFT) e características espaço-temporais (usando o descritor STIP (*Space-Time Interest Points*) [Laptev, 2005]). Assim como em [Avila et al., 2013], as características locais e de cor são extraídas dos quadros centrais de cada tomada de vídeo. A característica espaço-temporal é extraída de cada tomada de vídeo. Todas as características foram codificadas utilizando o modelo BoW e para classificação o classificador SVM. Em uma análise isolada de cada característica, Valle et al. obtêm os melhores resultados com características espaço-temporais. É também proposto um esquema de classificação em que o rótulo final da classificação é obtido através de uma votação majoritária sobre a opinião dos classificadores utilizados para cada característica separada.

Um estudo sobre o impacto de padrões de movimento baseados em classificação de ações foi feito por Souza et al. [2012]. Informações de cor foram incorporadas no descritor STIP para detecção e descrição de padrões de movimento aplicadas no contexto de detecção de violência e pornografia em vídeos. Desta maneira, três extensões do descritor STIP foram apresentadas: ColorSTIP, usando um modelo de cor RGB normalizado; HueSTIP, baseado em uma saturação ponderada do canal de tonalidade (Hue); e Hue-ColorSTIP, uma combinação entre as duas abordagens anteriores. Modelos BoW foram aplicados para codificar as características espaço-temporais e SVM como classificador final.

Partindo do pressuposto que materiais pornográficos podem apresentar movimentos e padrões característicos de áudio, Rea et al. [2006] propõem o uso de informações multimodal extraídas do fluxo audiovisual de vídeos. Os autores usam informações de movimento, diretamente extraídas de vetores de movimento MPEG, para melhorar a segmentação de regiões com informação de pele. Além disso, foi apresentada uma característica de áudio para detectar materiais de áudio pornográfico, baseada principalmente na ocorrência de padrões periódicos na "energia" do fluxo de áudio. Tanto a segmentação de regiões de pele, baseada em cor e movimento, quanto a extração da característica de áudio são realizadas em tempo real.

Zuo et al. [2008] propõem um *framework* para reconhecimento de filmes pornográficos baseado na fusão de informações de áudio e vídeo. Como características de áudio, são utilizadas MFCC (*Mel-Frequency Cepstral Coefficients*) [Sahidullah & Saha, 2012]). Um classificador GMM (*Gaussian Mixture Model*) é utilizado para reconhecer

sons pornográficos. Um algoritmo baseado em forma é utilizado para detectar os quadros com conteúdos pornográficos de um vídeo. Por fim, é utilizado um algoritmo de fusão baseado em teoria Bayesiana para combinar os resultados de reconhecimento de áudio e vídeo.

Liu et al. [2011] combinam informações de áudio com métodos baseados em *color moments* e histogramas de bordas. Diversas características de áudio, como MFCC, são extraídas em intervalos do vídeo baseados em padrões periódicos. Em seguida, o vídeo é descrito utilizando o modelo BoW para codificar as características de áudio. Dois classificadores SVM são utilizados para classificação (um para o modelo BoW de áudio e outro para os métodos baseados *color moments* e histogramas de bordas). Por fim, uma fusão tardia de classificadores, baseada em uma ponderação dos *scores*, é utiliza para dar a classificação final do vídeo.

De forma semelhante, Ulges et al. [2012] realizam uma segmentação regular do vídeo em janelas de tempo de 4 segundos. Para cada uma destas janelas de tempo, são extraídas características multimodais a partir do conteúdo audiovisual: (i) histogramas de movimento, a partir de vetores de movimento MPEG; (ii) características de cor relacionadas à pele, como em [Jones & Rehg, 2002]; (iii) características locais, como em [Deselaers et al., 2008]; e (iv) características MFCC, extraídas a partir do sinal de áudio. Para cada característica extraída, um classificador SVM é utilizado e, por fim, a predição do vídeo é dada com base em uma fusão tardia baseada nos *scores* de cada classificador.

Capítulo 4

Metodologia Proposta

Neste capítulo, é descrita a metodologia aplicada no presente trabalho para detecção de pornografia em vídeos. O capítulo está organizado da seguinte forma: na Seção 4.1, é descrita a abordagem base utilizada para detecção de pornografia em vídeos e cada um dos seus passos é detalhado nas subseções seguintes; na Seção 4.2, é apresentado o descritor de vídeo proposto usado no contexto de detecção de pornografia, através da aplicação da função mediana em vetores da representação intermediária BossaNova.

4.1 Abordagem Base - Classificação Utilizando Descritores Binários

Na Figura 4.1, é apresentado o esquema geral utilizado para detecção de pornografia em vídeos. Para isto, é utilizado o mesmo método proposto por Avila et al. [2013], porém se diferenciando nas etapas de extração de características e criação do dicionário visual.

Inicialmente, são extraídos quadros-chave do vídeo. Em seguida, são extraídas as características locais destes quadros-chave, que serão utilizadas para a descrição do conteúdo de cada quadro-chave. A seguir, na fase de treinamento, é criado o dicionário visual baseado em uma amostragem das características locais extraídas. Feito isso, as características locais são codificadas por uma representação intermediária e então um classificador é treinado. Na fase de classificação, a opinião do classificador é solicitada para cada quadro extraído do vídeo, e, a decisão final, é realizada por meio de uma votação majoritária. Cada um destes passos é detalhado a seguir.

Seleção de Quadros-Chave do Vídeo

Segundo Truong & Venkatesh [2007], quadros-chave são definidos como um con-

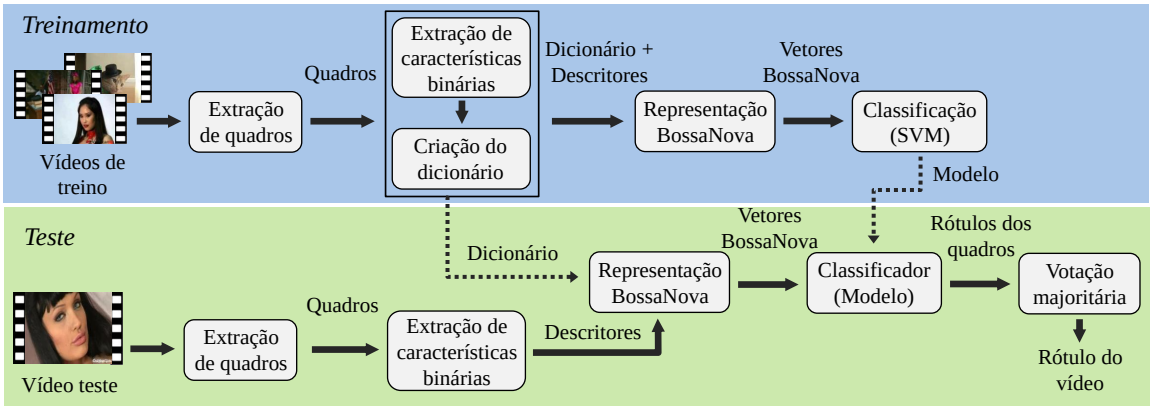


Figura 4.1. Visão geral da abordagem de detecção de pornografia proposta utilizando descritores binários e representação BossaNova.

junto de imagens marcantes extraídas da fonte do vídeo base. Existem várias formas de se obter os quadros-chave de um vídeo, entre elas, uma abordagem, largamente utilizada em tarefas que envolvam análise de vídeos, é baseada na segmentação do vídeo em tomadas e assim extrair um quadro de cada tomada (primeiro, último ou central) que possa representar o conteúdo daquela tomada.

Embora existam formas sofisticadas para realizar a seleção de quadros-chave para representar o vídeo, assim como em [Avila et al., 2013], este trabalho optou em realizar esta tarefa de uma forma relativamente simples, selecionado os quadros centrais de cada tomada presente no vídeo.

Extração de Característica Locais

Apesar do fato de que descritores de características locais, como SIFT ou SURF, apresentem boa precisão quando usados em abordagens baseadas nos modelos *Bag-of-Words* (BoW), estes apresentam um alto custo computacional no cálculo de seus vetores de características gerando vetores de alta dimensionalidade compostos por valores reais, tornando-se extremamente difícil de usá-los em aplicações em tempo real ou que necessitem de um baixo custo de processamento.

Como uma alternativa de baixa complexidade, os descritores binários têm emergido recentemente. Este tipo de descritor tem recebido uma atenção considerável por gerar resultados similares, ou até mesmo melhores, quando comparados aos descritores não-binários do estado da arte. Desta maneira, para a etapa de extração de características, foram utilizados descritores binários extraídos de maneira densa (amostragem densa).

Criação do Dicionário Visual

Com as características locais em mãos, estas devem ser codificadas por meio de uma representação intermediária, para serem usadas na etapa de classificação. No entanto, um dicionário visual deve ser criado antes da etapa de codificação.

Para a construção do dicionário visual, em vez de aplicar o clássico algoritmo de agrupamento *k-means*, foi utilizado o algoritmo *k-medians* [Jain & Dubes, 1988]. A razão principal dessa escolha é por causa das características empregadas. Como os descritores binários são representados por valores binários, o agrupamento utilizando *k-means* resultaria em valores não-binários. Além disso, segundo Bradley et al. [1997], o algoritmo *k-medians* produz um agrupamento melhor que o clássico *k-means* pelo fato de *outliers* apresentarem menos influência sobre o algoritmo. Ainda, a distância Euclidiana é substituída pela distância de Hamming, a fim de calcular a distância entre os descritores e os centróides.

Meio de Representação Intermediário

Em relação à representação intermediária, o modelo BoW, e suas extensões, é o meio de representação intermediário mais comum utilizado em abordagens de detecção de pornografia para codificar as características locais extraídas de imagens (ver Seção 3.2).

Devido aos bons resultados encontrados na literatura entre a combinação de descritores binários e modelos BoW para reconhecimento visual, este trabalho propõe utilizar o recente meio de representação intermediário BossaNova, que enriquece a representação BoW mantendo um histograma das distâncias entre os descritores encontrados na imagem e cada elemento do dicionário visual.

Classificação

Para a etapa de classificação foi utilizado o algoritmo *Support Vector Machines* (SVM). A escolha deste algoritmo dá-se pelo fato que diversos trabalhos encontrados na literatura que envolvem o reconhecimento visual, como detecção de pornografia, utilizam este classificador demonstrando excelentes resultados e até mesmo superando outros meios de classificação quando empregados a tarefas de reconhecimento visual [Deselaers et al., 2008; Zhou et al., 2010; Lee et al., 2013; Avila et al., 2013; Ries & Lienhart, 2014].

Para a fase de treinamento, o SVM é aplicado como classificador atuando sobre os vetores BossaNova de cada quadro-chave extraído do vídeo, rotulando-os como quadros pornográficos ou quadros não-pornográficos. Durante a fase de teste, também, é aplicado o classificador SVM, porém a predição final do vídeo é dada por meio de uma

votação majoritária sobre a predição dos quadros-chave feita pelo SVM.

4.2 Descritor de Vídeo Proposto - BossaNova Video Descriptor

Como pode ser visto em [Caetano et al., 2014], é possível utilizar descritores binários como características locais em conjunto com a representação BossaNova para tarefas de reconhecimento visual, especificamente detecção de pornografia. Apesar dos resultados promissores para a detecção de pornografia, a classificação final do vídeo é obtida por uma votação majoritária sobre as imagens. Uma desvantagem é que o número de quadros-chave pornográficos deve ser maior do que o número de quadros-chave não-pornográficos, que nem sempre é verdade. Além disso, a abordagem com votação majoritária leva em consideração apenas imagens estáticas, ignorando a probabilidade de classificações corretas e erradas.

Com o objetivo de resolver estes problemas, em vez de realizar a predição final do vídeo por meio de uma votação majoritária, é apresentado um novo descritor de vídeo que melhora o desempenho da classificação combinando as "assinaturas" dos quadros-chave da seguinte forma.

Considere $\mathcal{V}$ como uma sequência de vídeo. $\mathcal{V} = \{f^i\}$, $i \in [1, N]$, onde f^i é o quadro central selecionado da tomada i e N é o número de tomadas. Considere $\mathcal{Z}$ como um conjunto de vetores BossaNova calculados para o vídeo $\mathcal{V}$. $\mathcal{Z} = \{\mathbf{z}^i\}$, $i \in [1, N]$, onde $\mathbf{z}^i$ é o vetor BossaNova extraído para o quadro f^i. O descritor de vídeo $\mathcal{Z}$ pode ser modelado pela função h da seguinte maneira:

$$\begin{aligned} h : \mathbb{R}^Z &\longrightarrow \mathbb{R}^Z, \\ \mathcal{Z} &\longrightarrow h(\{\mathbf{z}^i\}) = \left[\left[o_{m,b}\right], p_m\right]^{\mathrm{T}}, \\ o_{m,b} &= \text{mediana}(z^i_{m,b}), \\ p_m &= \text{mediana}(t^i_m), \end{aligned} \tag{4.1}$$

onde $Z \subset \{1, ..., M\} \times \{1, ..., B\}$, e $\mathbf{z}^i = \left[\left[z^i_{m,b}\right], t^i_m\right]^{\mathrm{T}}$.

Intuitivamente, este novo descritor de vídeo representa a distância mediana para cada característica local em relação à palavra visual no dicionário visual, uma vez que cada representação BossaNova contém informações sobre a distribuição de distâncias da palavra visual (histograma de distâncias). Além disso, valores *outliers* são eliminados pela função mediana.

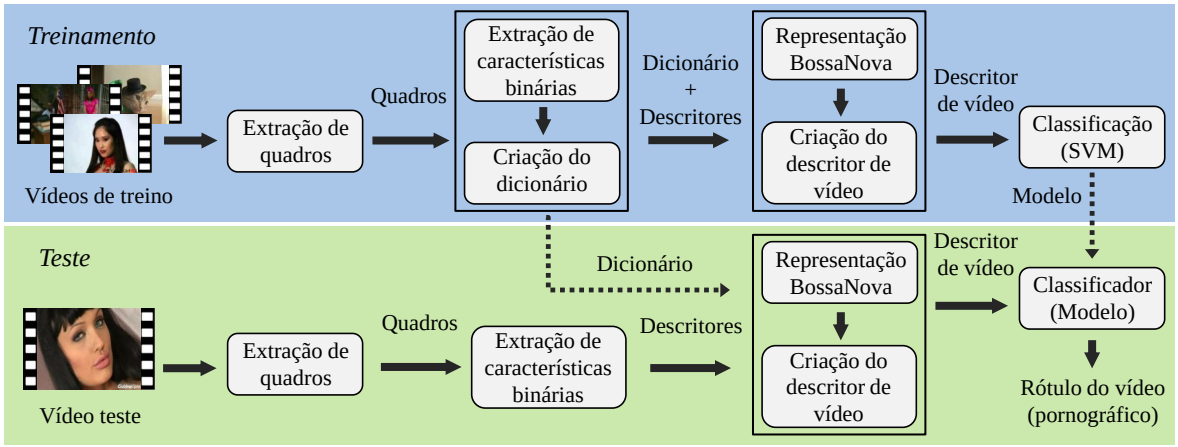


Figura 4.2. Visão geral da abordagem para detecção de pornografia utilizando o descritor de vídeo proposto, *BossaNova Video Descriptor* (BossaNova VD).

Na Figura 4.2, é ilustrada a visão geral da abordagem proposta para detecção de pornografia utilizando o descritor de vídeo. É importante notar que esta metodologia é uma adaptação de Avila et al. [2013], em que as características locais da imagem são extraídas seguidas pela codificação com a representação BossaNova para cada quadro-chave, porém, como na Seção 4.1, são utilizados descritores binários para a etapa de extração de características e criação do dicionário visual com *k-medians* e distância de Hamming. O descritor de vídeo proposto é calculado de acordo com a Equação 4.1 para representar cada vídeo a ser classificado e então um classificador SVM é utilizado para dar a predição final do vídeo.

Capítulo 5

Resultados Experimentais

Este capítulo se concentra na descrição das bases de dados utilizadas neste trabalho bem como dos critérios de avaliação, parâmetros e experimentos realizados. O capítulo está organizado da seguinte forma: na Seção 5.1, são descritas as bases de dados utilizadas, PASCAL VOC 2007 [Everingham et al., 2007] e *Pornography* [Avila et al., 2011, 2013]; na Seção 5.2, é apresentado o critério de avaliação utilizado para medir o desempenho da classificação, assim como o protocolo de classificação utilizado; e, por fim, na Seção 5.3, são mostrados os experimentos realizados e são analisados os resultados alcançados.

Todos os experimentos foram realizados em uma máquina Linux 64 *bits* (Ubuntu 12.04) com processador Intel® Xeon® CPU X5670 @ 2.93 GHz com 24 núcleos e 70 GB de RAM. Apesar do grande poder computacional disponível, este não foi totalmente utilizado para processar os resultados experimentais. Além disso, não foram utilizadas GPUs ou outras otimizações de hardware. O código fonte é escrito nas linguagens C, C++ e Java.

5.1 Bases de Dados

Para a realização dos experimentos, foram utilizadas duas bases de dados desafiadoras: PASCAL VOC 2007 [Everingham et al., 2007], usada para validar os descritores binários em conjunto com a representação intermediária BossaNova, em tarefas de classificação que envolvam reconhecimento visual; e *Pornography* [Avila et al., 2011, 2013], para o contexto de detecção de pornografia, aplicada para a abordagem base e para o descritor de vídeo proposto.

Figura 5.1. Exemplos extraídos da base de dados PASCAL VOC 2007 [Everingham et al., 2007].

5.1.1 PASCAL VOC 2007

O desafio PASCAL *Visual Object Classes* (VOC) é um *benchmark* em reconhecimento visual de categorias de objetos organizado anualmente (de 2005 até 2012), composto por duas competições principais: (i) classificação, para cada uma das classes, prever a presença/ausência de um exemplo de uma das classes de objeto na imagem de teste; e (ii) detecção, prover uma caixa delimitadora e um rótulo para cada objeto das classes de objeto encontrado na imagem de teste.

A base de dados PASCAL VOC 2007 [Everingham et al., 2007] é composta por fotografias rotuladas recolhidas do *site* de compartilhamento de fotos Flickr. O objetivo deste desafio é reconhecer 20 classes compostas por objetos em cenas realistas (*avião*, *bicicleta*, *pássaro*, *barco*, *garrafa*, *ônibus*, *carro*, *gato*, *cadeira*, *vaca*, *mesa de jantar*, *cachorro*, *cavalo*, *moto*, *pessoa*, *vaso de plantas*, *ovelha*, *sofá*, *trem*, *tv/monitor*). No total, a base de dados conta com 9963 imagens. Alguns exemplos são apresentados na Figura 5.1.

A base de dados PASCAL VOC 2007 é uma referência em classificação de imagens, contendo uma variabilidade significativa em termos de tamanho do objeto, orientação, pose, iluminação, posição e oclusão. A base de dados está disponível gratuitamente[1].

5.1.2 *Pornography*

Devido às questões de direitos autorais e potenciais limitações legais sobre a distribuição de grandes quantidades de material pornográfico, os trabalhos de detecção de pornografia encontrados na literatura geralmente utilizam suas próprias bases de dados, tornando difícil a comparação dos resultados obtidos entre trabalhos.

Avila et al. [2011, 2013] criaram uma base de dados representativa composta por vídeos da Internet, tanto pornográficos quanto não-pornográficos. Para a classe de pornografia, os vídeos foram extraídos de *sites* fornecedores deste tipo de material[2]. Para a classe não-pornográfica, foram extraídos vídeos de propósito geral do YouTube[3].

A base de dados *Pornography* [Avila et al., 2013] contém cerca de 80 horas de vídeos, sendo 400 vídeos pornográficos e 400 não-pornográficos. A classe pornográfica é composta por vários gêneros de pornografia e pessoas de diferentes etnias, incluindo os multi-étnicos. A classe não-pornográfica é dividida em duas sub-classes: (i) *"fácil"*, com 200 vídeos selecionados aleatoriamente da Internet; e (ii) *"difícil"*, com 200 vídeos selecionados a partir de consultas de busca textuais como "praia", "luta" e "natação", sendo um grupo de vídeos com bastante exposição de pele tornando a base de dados altamente desafiadora. Além disso, a base de dados *Pornography* disponibiliza os vídeos segmentados em tomadas (um total de 16.727 tomadas de vídeo). Na Tabela 5.1, é possível encontrar um resumo da base.

A Figura 5.2 apresenta alguns quadros selecionados a partir de uma pequena amostra da base de dados, ilustrando a diversidade entre os vídeos pornográficos e os

Tabela 5.1. Resumo da base de dados *Pornography*.

Classe	Vídeos	Horas	Tomadas por vídeo
Pornográfica	400	57,0	15,6
Não-pornográfica (*"fácil"*)	200	11,5	33,8
Não-pornográfica (*"difícil"*)	200	8,5	17,5
Todos os vídeos	**800**	**77,0**	**20,6**

[1] `http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2007/`
[2] Por exemplo, http://www.{RedTube, XTube, PornTube, Xvideos}.com
[3] `http://www.youtube.com`

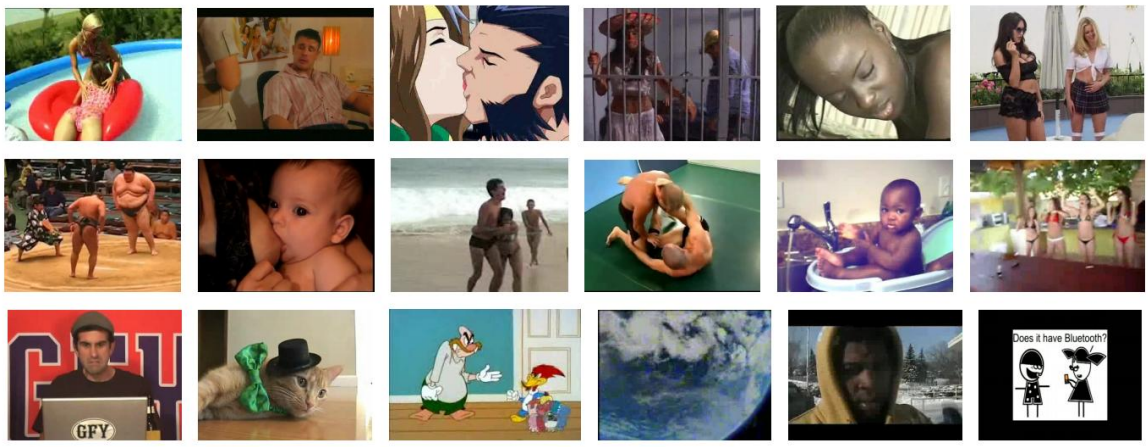


Figura 5.2. Quadros selecionados a partir de uma amostra da base de dados *Pornography* [Avila et al., 2013]. Na linha superior é ilustrado a diversidade contida na classe pornográfica, na linha central é exibida a sub-classe não-pornográfica *"difícil"* e na última linha a sub-classe não-pornográfica *"fácil"*.

vídeos não-pornográficos *"difícil"* e *"fácil"*. A base de dados não está disponível para a comunidade em geral, devido aos problemas de direitos autorais, mas o acesso a ela pode ser concedido via solicitação[4].

5.2 Critérios de Avaliação

Nesta seção, primeiramente, são definidos os protocolos de treinamento/teste utilizados em cada base de dados nos experimentos e, em seguida, são discutidas as medidas de avaliação usadas para cada base de dados.

5.2.1 Protocolo de Validação

PASCAL VOC 2007

O desafio VOC 2007 é composto por duas competições principais: classificação e detecção. Para os experimentos realizados neste trabalho, são apresentados apenas os resultados para a tarefa de classificação. O conjunto de dados é dividido em três subgrupos: (i) treinamento, composto por 2501 imagens; (ii) validação, composto por 2510 imagens; e (iii) teste, composto por 4952 imagens. Os resultados experimentais são obtidos nos grupos treinamento+validação/teste.

A Tabela 5.2 resume o número de imagens para cada classe nos conjuntos de treinamento, validação e teste.

[4] `http://www.npdi.dcc.ufmg.br/pornography`

Tabela 5.2. Número de imagens para cada classe na base de dados PASCAL VOC 2007 [Everingham et al., 2007].

Classe	**#treinamento**	**#validação**	**#teste**
1: avião	112	126	204
2: bicicleta	116	127	239
3: pássaro	180	150	282
4: barco	81	100	172
5: garrafa	139	105	212
6: ônibus	97	89	174
7: carro	376	337	721
8: gato	163	174	322
9: cadeira	224	221	417
10: vaca	69	72	127
11: mesa de jantar	97	103	190
12: cachorro	203	218	418
13: cavalo	139	148	274
14: moto	120	125	222
15: pessoa	1025	983	2007
16: vaso de plantas	133	112	224
17: ovelha	48	48	97
18: sofá	111	118	223
19: trem	127	134	259
20: tv/monitor	128	128	229

Para a classificação, é aplicada a estratégia um-contra-todos (*one-versus-all*), que consiste em comparar cada uma das n classes contra todas as outras. Um problema multi-classe com n classes é decomposto em n problemas binários: uma classe principal contra as classes restantes.

Pornography

Com o objetivo de estabelecer uma comparação justa, neste trabalho foi utilizado o mesmo protocolo de validação usado por Avila et al. [2011, 2013]. O protocolo experimental aplicado foi a validação cruzada com 5 *folds* (5-*fold cross-validation*), gerando cerca de 640 vídeos para treinamento e 160 para o teste em cada *fold*. O número de tomadas para cada conjunto de treinamento e teste é apresentado na Tabela 5.3.

Tabela 5.3. Número de tomadas para cada conjunto de treinamento e teste para base de dados *Pornography* [Avila et al., 2013]. No total, cada execução contém cerca de 640 vídeos para treinamento e 160 para teste.

Execuções	**#treinamento**		**#teste**	
	não-pornografia	pornografia	não-pornografia	pornografia
execução1	8194	4909	2146	1478
execução2	8488	4933	1852	1454
execução3	8470	5144	1870	1243
execução4	8351	5262	1989	1125
execução5	7857	5300	2483	1087

5.2.2 Medidas de Avaliação

PASCAL VOC 2007

No desafio PASCAL VOC 2007, a principal medida de avaliação utilizada para a tarefa de classificação é o *mean Average Precision* (mAP). O *Average Precision* (AP) leva em conta tanto a precisão quanto a revocação, medindo a qualidade da classificação das imagens. O AP é calculado usando a Equação 5.1:

$$AP = \frac{\sum_{n=1}^{N} Pr(n) \times rel(n)}{r_i}, \tag{5.1}$$

onde r_i é o número total de imagens relevantes para a classificação, N é o número de imagens, $Pr(n)$ é a precisão da classificação da imagem n e $rel(n)$ é uma função binária sobre a relevância da n-ésima imagem (Equação 5.2).

$$rel(n) = \begin{cases} 1, & \text{se a } n\text{-ésima imagem for relevante} \\ 0, & \text{caso contrário} \end{cases}$$

Assim sendo, o valor de mAP é a média dos valores AP para todas as classes de imagens.

Além disso, para verificar se há diferença estatística entre os resultados, foi realizado o teste de observações pareadas, conhecido como *paired t-test* [Jain, 1991]. Foi usado um nível de confiança igual a 95%, onde as observações provenientes do AP, por classe, foram utilizadas para o *paired t-test*. Este teste é usado para determinar se duas abordagens são estatiticamente significantes uma da outra.

Pornography

Para a base de dados *Pornography*, foram utilizadas duas medidas de avaliação comuns aplicadas por trabalhos que utilizaram esta base de dados. Essas medidas são definidas a seguir:

- Matriz de Confusão: é uma tabela onde cada coluna representa as amostras preditas em uma classe, enquanto que a linha representa a classe real. Cada elemento T_{ij} da matriz representa o número de amostras classificados como classe i e referenciados como classe j. Essa tabela permite uma fácil visualização do número de amostras classificadas corretamente e erroneamente.

- Acc (Acurácia): é a porcentagem de amostras classificadas corretamente (ver Equação 5.2):

$$Acc = \frac{\sum_{i}^{K} T_{ii}}{\sum_{ij}^{K} T_{ij}} \times 100 \tag{5.2}$$

5.3 Experimentos

Nos experimentos, é investigado o poder dos descritores binários em conjunto com a recente representação intermediária BossaNova para tarefas de classificação. Os objetivos são: (i) validar os descritores binários para tarefas de classificação em reconhecimento visual; (ii) comparar a performance dos descritores binários para tarefas de classificação no contexto de detecção de pornografia; e (iii) avaliar o descritor de vídeo proposto também no contexto de detecção de pornografia.

A fim de estudar o comportamento dos descritores binários, em conjunto com a representação intermediária BossaNova, para ambas as bases de dados, cinco descritores binários (BRIEF [Calonder et al., 2010], ORB [Rublee et al., 2011], BRISK [Leutenegger et al., 2011], FREAK [Ortiz, 2012] e BinBoost [Trzcinski et al., 2013]) são extraídos de maneira densa a cada 6 *pixels* (amostragem densa). Os códigos dos descritores binários foram obtidos, exceto BinBoost[5], a partir do repositório OpenCV [Bradski, 2000], uma das bibliotecas mais populares de visão computacional. Todos os descritores binários foram extraídos com seus parâmetros *default*. Para a representação

[5]A implementação do descritor BinBoost está disponível em `http://www.cvlab.epfl.ch/research/detect/binboost`.

intermediária BossaNova, foi utilizado o código disponibilizado[6]. Além disso, para o classificador SVM, foi utilizada a implementação disponível pela biblioteca JKernelMachines [Picard et al., 2013].

Para criação do dicionário visual, foi aplicado o algoritmo de clusterização *k-medians* com a distância de Hamming sobre uma amostra aleatória de 500 mil descritores, para a base de dados PASCAL VOC 2007, e um milhão de descritores para a base de dados *Pornography*.

Para todos os experimentos realizados, um *kernel* não-linear Gauss-ℓ_2 foi utilizado para o *Support Vector Machines* (SVM). A matriz de distâncias do *kernel* é calculada como $exp(-\gamma d(x, x'))$, com d sendo a distância e γ sendo definido como o inverso da distância média entre os pares.

Com o propósito de uma comparação mais justa, as comparações dos resultados na base de dados PASCAL VOC 2007 são feitas apenas com trabalhos que também empregaram descritores binários. Para a base de dados *Pornography*, os resultados foram comparados com Avila et al. [2011, 2013] e com o software PornSeer Pro, um sistema de detecção de pornografia da indústria baseado na detecção de características específicas (como seios e órgãos sexuais) examinando cada quadro individual do vídeo.

Nesta seção, primeiramente, são apresentados os experimentos de validação dos descritores binários realizados na PASCAL VOC 2007 (Seção 5.3.1). Em seguida, são apresentados os experimentos realizados com a abordagem base no contexto de detecção de pornografia (Seção 5.3.1). Por fim, são apresentados os experimentos para o descritor de vídeo proposto (Seção 5.3.3).

5.3.1 Resultados da Validação dos Descritores Binários

Um fluxograma do esquema usado para os experimentos na base de dados PASCAL VOC 2007 é descrito na Figura 5.3.

Conforme descrito na Seção 2.2.2, a representação intermediária BossaNova é composta por três parâmetros principais: o intervalo de distâncias $[\alpha_m^{min}, \alpha_m^{max}]$, o número de bins do histograma local B e o tamanho do dicionário visual M). Para estes experimentos, os parâmetros foram mantidos como em Avila et al. [2013]. O intervalo de distâncias $[\alpha_m^{min}, \alpha_m^{max}]$ foi definido como $\alpha_m^{min} = \lambda_{min} \cdot \sigma_m$ e $\alpha_m^{max} = \lambda_{max} \cdot \sigma_m$, onde σ_m é o desvio padrão de cada *cluster* $\mathbf{c}_m$ obtido pelo algoritmo de clusterização *k-medians*. Assim sendo, os demais parâmetros foram definidos como: $B = 2$, $\lambda_{min} = 0{,}4$ e $\lambda_{max} = 2{,}0$, e $M = 1024$.

[6] `http://www.npdi.dcc.ufmg.br/bossanova`

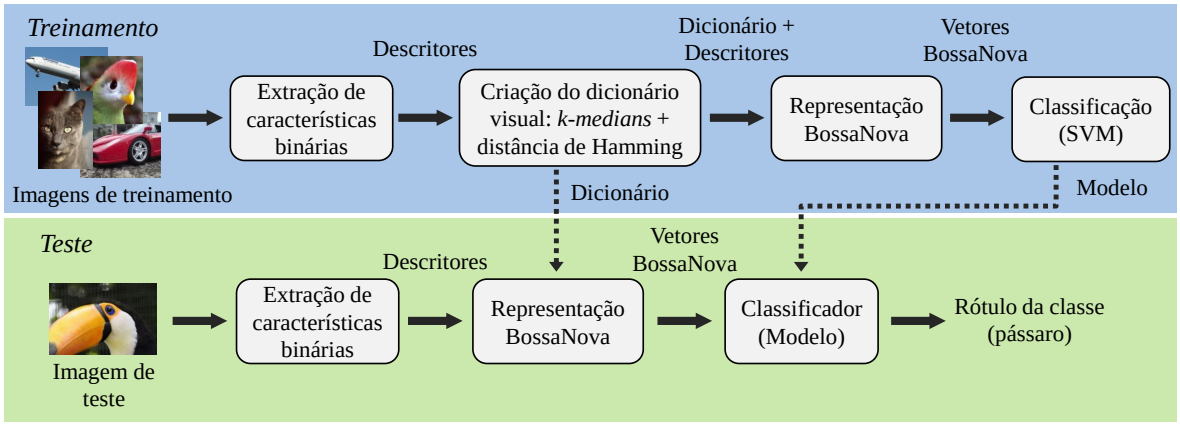


Figura 5.3. Fluxograma do esquema usado para os experimentos na base de dados PASCAL VOC 2007.

A Tabela 5.4 apresenta os resultados dos experimentos realizados na base de dados PASCAL VOC 2007. Pode-se notar que a abordagem proposta (descritor binário e representação BossaNova) supera os métodos anteriores, que utilizaram o modelo *Bag-of-Words* (BoW) clássico. Nos resultados alcançados, pode-se notar que o descritor BinBoost fornece o melhor resultado (mAP = 43,20%), enquanto FREAK apresenta o menor resultado (mAP = 33,32%).

A Tabela 5.4 também apresenta uma comparação com os resultados da literatura. Em comparação com Zhang et al. [2013], todos os resultados alcançados pela abordagem proposta demostram ser superiores ao método de Zhang et al. utilizando LBP (*Local Binary Pattern*) + BoW com uma escala (mAP = 33,24%). Também é importante ressaltar que Zhang et al. utilizaram um dicionário visual ligeiramente maior que o

Tabela 5.4. Resultados de classificação de imagens (mAP %) dos experimentos realizados e trabalhos da literatura para a base de dados PASCAL VOC 2007.

	Abordagem	**mAP (%)**
Resultados da literatura	BoW (LBP, uma escala) [Zhang et al., 2013]	33,24
	BoW (LBP, multi-escala) [Zhang et al., 2013]	35,17
	BoW (BRIEF) [Chatzilari et al., 2013]	21,54
	BoW (ORB) [Chatzilari et al., 2013]	21,62
Experimentos realizados	BossaNova (BRIEF)	36,22
	BossaNova (ORB)	37,14
	BossaNova (BRISK)	38,00
	BossaNova (FREAK)	33,32
	BossaNova (BinBoost)	**43,20**

utilizado pelos experimentos do presente trabalho (1200 palavras visuais contra 1024, respectivamente). O melhor resultado publicado de Zhang et al. é 35,17% para LBP + BoW (multi-escala). Em relação a este resultado, é possível observar que, mesmo não usando descritores multi-escala, os resultados alcançados com a abordagem proposta superam o resultado de Zhang et al., exceto para o descritor FREAK.

A comparação com Chatzilari et al. [2013] é particularmente relevante, porque, nos experimentos realizados, foram utilizados os mesmos descritores binários que eles (BRIEF e ORB com parâmetros *default*). É possível observar que os resultados alcançados, BossaNova + BRIEF (mAP = 36,22%) e BossaNova + ORB (mAP = 37,14%), são muito melhores do que os resultados de Chatzilari et al., BRIEF + BoW (mAP = 21,54%) e ORB + BoW (mAP = 21,62%). Além disso, os autores usaram um dicionário visual com tamanho de 2000 palavras visuais, enquanto os experimentos realizados neste trabalho utilizaram apenas 1024 palavras visuais. Em vista disso, os resultados alcançados pela abordagem proposta podem ser considerados muito bons, uma vez que dicionários visuais maiores resultam em melhores resultados [Chatfield et al., 2011].

Para verificar a significância estatística dos resultados obtidos, um teste estatístico para as diferenças entre as médias foi realizado utilizando *paired t-test*, pareado sobre as classes da base de dados. Na Tabela 5.5, é apresentado os testes entre a abordagem utilizando o descritor BinBoost e os demais descritores binários. O intervalo de confiança (IC) para a diferença média é calculado utilizando o modelo *Student t-test* e a diferença é considerada significativa se o intervalo não incluir zero (marcado com ✓). Para os testes desta seção, foi utilizada uma confiança de 95%.

Tabela 5.5. Teste estatístico *paired t-test*, com confiança de 95%, entre a abordagem utilizando o descritor BinBoost e os demais descritores binários. A diferença é considerada significativa se estiver marcada com ✓.

	Abordagem	**mAP %**	**CI (95%)**
1	BossaNova (BRIEF)	36,22	$1 \leftrightarrow 5$ ✓
2	BossaNova (ORB)	37,14	$2 \leftrightarrow 5$ ✓
3	BossaNova (BRISK)	38,00	$3 \leftrightarrow 5$ ✓
4	BossaNova (FREAK)	33,32	$4 \leftrightarrow 5$ ✓
5	BossaNova (BinBoost)	43,20	

Variação de Parâmetros - Descritor BinBoost

Devido aos excelentes resultados alcançados pelo descritor binário BinBoost na base de dados PASCAL VOC 2007, foram realizados experimentos variando o parâmetro d, relacionado ao tamanho do descritor, e o tamanho M do dicionário visual.

A Tabela 5.6 apresenta os resultados relacionados à variação do tamanho do descritor BinBoost (entre 8, 16 e 32 dimensões). Pode-se notar que o descritor BinBoost com 16 dimensões apresenta o melhor resultado (mAP = 44,6%), ultrapassando até mesmo sua variação com 32 dimensões. Também pode-se notar que, em um nível de confiança de 95%, a diferença é significativa para o descritor BinBoost com 16 dimensões quando comparado a 8 dimensões.

Tabela 5.6. Resultados de classificação de imagens (mAP %) relacionados à variação do tamanho d do descritor BinBoost na base de dados PASCAL VOC 2007.

	Abordagem	**mAP %**	**CI (95%)**
1	BossaNova (BinBoost $d = 8$)	43,2	
2	BossaNova (BinBoost $d = 16$)	**44,6**	$1 \leftrightarrow 2$ ✓
3	BossaNova (BinBoost $d = 32$)	44,2	$1 \leftrightarrow 3$, $2 \leftrightarrow 3$

A Tabela 5.7 apresenta os resultados relacionados a variação do tamanho do dicionário visual (1024, 2048, 4096 e 8192 palavras visuais) para a abordagem BossaNova + BinBoost ($d = 16$). É possível observar o impacto do tamanho do dicionário visual no desempenho da classificação, mostrando claramente que dicionários visuais maiores levam a um melhor resultado. Portanto, a abordagem proposta consegue alcançar um resultado de 46,8%, com um dicionário visual de tamanho $M = 8192$.

Tabela 5.7. Resultados de classificação de imagens (mAP %) relacionados à variação do tamanho M do dicionário visual para a abordagem BossaNova + BinBoost ($d = 16$) na base de dados PASCAL VOC 2007.

Dicionário	**mAP (%)**
$M = 1024$	44,6
$M = 2048$	45,5
$M = 4096$	46,2
$M = 8192$	**46,8**

5.3.2 Resultados da Abordagem Base - Classificação Utilizando Descritores Binários

Para estes experimentos, os parâmetros da representação BossaNova foram mantidos como em Avila et al. [2013]: $M = 256$, $B = 10$, $\lambda_{min} = 0$ e $\lambda_{max} = 3$.

A Tabela 5.8 apresenta os resultados alcançados com a abordagem base e os resultados reportados na literatura sobre a base de dados *Pornography*. Mais uma vez, é possível observar que, nos resultados alcançados, o descritor BinBoost com dimensionalidade $d = 16$ fornece o melhor resultado (89,40%). Além disso, é possível observar que o melhor resultado alcançado com a abordagem base está bem próximo ao melhor resultado relatado pela literatura. É importante notar que, o melhor resultado publicado é obtido utilizando descritores HueSIFT, uma variação do descritor SIFT incluindo informação de cor, o que é particularmente relevante para esta base de dados. Além disso, é possível observar a vantagem da abordagem base proposta (BossaNova + descritores binários) quando comparada com a abordagem BoW clássica, que também empregou descritores HueSIFT.

Tabela 5.8. Resultados de classificação de vídeos (Acc % e desvio-padrão) da abordagem base proposta e os resultados publicados sobre a base de dados *Pornography*.

	Abordagem	**Acc. (%)**
Resultados da literatura	BoW (HueSIFT) [Avila et al., 2013]	83,0 ± 3
	BossaNova (HueSIFT) [Avila et al., 2013]	**89,5** ± 1
Experimentos realizados	BossaNova (BRIEF)	86,03 ± 3
	BossaNova (ORB)	86,79 ± 3
	BossaNova (BRISK)	88,65 ± 2
	BossaNova (FREAK)	86,90 ± 3
	BossaNova (BinBoost $d = 8$)	87,28 ± 2
	BossaNova (BinBoost $d = 16$)	**89,40** ± 2
	BossaNova (BinBoost $d = 32$)	89,02 ± 2

As Tabelas 5.9, 5.10 e 5.11 apresentam as matrizes de confusão para a abordagem base proposta com BossaNova + BinBoost ($d = 16$), Avila et al. [2013] e o software PornSeer Pro. Como pode ser visto nas Tabelas 5.9 e 5.10, pode-se afirmar novamente que os resultados da abordagem base proposta com descritores binários está bem próximo do melhor resultado publicado [Avila et al., 2013], no entanto é possível notar que a abordagem proposta possui melhores resultados na classificação de vídeos

pornográficos enquanto que Avila et al. [2013] apresenta melhores resultados em vídeos não pornográficos.

A Tabela 5.12 apresenta uma comparação de tempo em relação ao: (i) tempo médio de extração dos descritores, (ii) tempo para criação do dicionário visual, e (iii) tempo médio para criar a representação BossaNova. É possível observar quão mais rápido os descritores binários são em relação ao descritor HueSIFT utilizado por Avila et al. [2013]. Os descritores BRIEF e ORB chegam a ter uma extração média 10 vezes mais rápida que seu concorrente não-binário e o descritor BinBoost, com dimensionalidade d = 32, chega a ser duas vezes mais rápido.

Tabela 5.9. Matriz de confusão para a abordagem de classificação base, BossaNova + BinBoost ($d = 16$).

		Vídeo rotulado como	
		pornográfico	não-pornográfico
Classe	pornográfico	90,5%	9,5%
do vídeo	não-pornográfico	11,7%	88,3%

Tabela 5.10. Matriz de confusão dos resultados de Avila et al. [2013].

		Vídeo rotulado como	
		pornográfico	não-pornográfico
Classe	pornográfico	88,2%	11,8%
do vídeo	não-pornográfico	9,2%	90,8%

Tabela 5.11. Matriz de confusão dos resultados do software PornSeer Pro.

		Vídeo rotulado como	
		pornográfico	não-pornográfico
Classe	pornográfico	65,1%	34,9%
do vídeo	não-pornográfico	12,5%	87,5%

É possível observar também na Tabela 5.12, os tempos de criação do dicionário visual para cada abordagem. Nesta avaliação, a abordagem de Avila et al. [2013] possui o melhor tempo. O motivo disto se dá pelo fato do presente trabalho utilizar o algoritmo *k-medians*, enquanto Avila et al. utilizou o algoritmo *k-means*. No entanto, é importante ressaltar que a etapa de criação do dicionário visual é realizada durante

a fase de treinamento, ou seja, uma fase *offline*.

Tabela 5.12. Comparação de tempo (em segundos) em relação ao: (i) tempo médio de extração dos descritores, (ii) tempo para criação do dicionário visual, e (iii) tempo médio para criar a representação intermediária BossaNova.

Abordagem	**Descritor**	**Dicionário**	**BossaNova**
BossaNova (HueSIFT) [Avila et al., 2013]	$2,54$	$\mathbf{1,61 \times 10^3}$	$1,33$
BossaNova (BRIEF)	$0,24$	$4,75 \times 10^3$	$0,50$
BossaNova (ORB)	$\mathbf{0,23}$	$6,51 \times 10^3$	$0,49$
BossaNova (BRISK)	$0,64$	$10,23 \times 10^3$	$0,83$
BossaNova (FREAK)	$0,31$	$9,48 \times 10^3$	$0,84$
BossaNova (BinBoost $d = 8$)	$0,69$	$3,14 \times 10^3$	$\mathbf{0,12}$
BossaNova (BinBoost $d = 16$)	$0,81$	$16,12 \times 10^3$	$0,23$
BossaNova (BinBoost $d = 32$)	$1,02$	$12,73 \times 10^3$	$0,51$

Combinação de Classificadores

Devido aos bons resultados alcançados com a abordagem base proposta, este trabalho também realizou experimentos relacionados às combinações dos classificadores finais utilizados por cada descritor binário na abordagem base. O principal objetivo em combinar múltiplos métodos de classificação é produzir uma decisão final que seja melhor que uma única decisão [Kuncheva, 2004].

Para este propósito, foram realizados experimentos utilizando um *framework* de combinação de classificadores proposto por Faria et al. [2014]. Este *framework* combina automaticamente os classificadores mais discriminativos usando SVM. Para isto, são utilizadas medidas de diversidade para selecionar os classificadores menos correlacionados, porém eficazes.

O *framework* de Faria et al. faz uso de cinco medidas de diversidade: *Correlation Coefficient* (COR), *Double-Fault Measure* (DFM), *Disagreement Measure* (DM), *Interrater Agreement k* (IA), e *Q-Statistic* (QSTAT) definidas a seguir:

$$COR(C_i, C_j) = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{5.3}$$

$$DFM(C_i, C_j) = d \tag{5.4}$$

$$DM(C_i, C_j) = \frac{b+c}{a+b+c+d} \tag{5.5}$$

$$IA(C_i, C_j) = \frac{2(ac-bd)}{(a+b)(c+d)+(a+c)(b+d)} \tag{5.6}$$

$$QSTAT(C_i, C_j) = \frac{ad-bc}{ad+bc} \tag{5.7}$$

onde a é a porcentagem de vídeos em que ambos classificadores C_i e C_j classificaram corretamente na fase de treinamento, b e c são as porcentagens de vídeos que C_j classificou corretamente e C_i erroneamente e vice versa, e d é a porcentagem de vídeos que ambos classificadores classificaram incorretamente.

Faria et al. utilizaram uma estratégia baseada no paradigma de "aprendizagem combinada" [Rokach, 2010], cujo princípio é combinar classificadores fracos com o objetivo de construir um classificador global mais eficaz e eficiente. Baseando-se nesta ideia, o presente trabalho realizou experimentos com classificadores enfraquecidos diminuindo o parâmetro de custo C do SVM. Para grandes valores de C, o algoritmo escolherá um hiperplano com margem menor tentando fazer que todos os pontos de treinamento sejam classificados corretamente. Por outro lado, um valor muito pequeno de C fará com que o algoritmo procure por um hiperplano de margem maior, mesmo que este hiperplano classifique incorretamente alguns pontos.

Além disso, para utilizar o *framework*, foi necessário dividir o conjunto de treinamento da base de dados *Pornography* em dois sub-conjuntos: treinamento e validação. Para isto, foram utilizados os mesmos *folds* utilizados nos experimentos anteriores, porém divididos em treinamento, validação e teste. A Tabela 5.13 exibe o protocolo utilizado.

As Tabelas 5.14 e 5.15 apresentam os resultados alcançados sobre o sub-conjunto de validação para todos os descritores binários testados e classificadores SVM $C = 0,1$ e $C = 1$.

A Tabela 5.16 apresenta os resultados obtidos, sobre o conjunto de teste, com o uso do *framework* para combinação entre os classificadores utilizados nos experimentos apresentados anteriormente. É possível observar um pequeno ganho nos *folds* 1, 2 e 3, porém, no resultado final (Média), a acurácia obtida se mostra menor que o melhor resultado apresentado na Tabela 5.15 (BinBoost $d = 32$) demonstrando que o descritor

Tabela 5.13. Protocolo utilizado para o *framework* de combinação de classificadores [Faria et al., 2014].

Execuções	**treinamento**	**validação**	**teste**
execução1	*fold* 1, 2 e 3	*fold* 4	*fold* 5
	fold 1, 2 e 4	*fold* 3	*fold* 5
	fold 1, 3 e 4	*fold* 2	*fold* 5
	fold 2, 3 e 4	*fold* 1	*fold* 5
execução2	*fold* 1, 2 e 3	*fold* 5	*fold* 4
	fold 1, 2 e 5	*fold* 3	*fold* 4
	fold 1, 3 e 5	*fold* 2	*fold* 4
	fold 2, 3 e 5	*fold* 1	*fold* 4
execução3	*fold* 1, 2 e 4	*fold* 5	*fold* 3
	fold 1, 2 e 5	*fold* 4	*fold* 3
	fold 1, 4 e 5	*fold* 2	*fold* 3
	fold 2, 4 e 5	*fold* 1	*fold* 3
execução4	*fold* 1, 3 e 4	*fold* 5	*fold* 2
	fold 1, 3 e 5	*fold* 4	*fold* 2
	fold 1, 4 e 5	*fold* 3	*fold* 2
	fold 3, 4 e 5	*fold* 1	*fold* 2
execução5	*fold* 2, 3 e 4	*fold* 5	*fold* 1
	fold 2, 3 e 5	*fold* 4	*fold* 1
	fold 2, 4 e 5	*fold* 3	*fold* 1
	fold 3, 4 e 5	*fold* 2	*fold* 1

BinBoost sozinho consegue atingir melhores resultados.

Tabela 5.14. Resultados de classificação (Acc % e desvio-padrão) sobre o subconjunto de validação, para a base de dados *Pornography*, com SVM $C = 0,1$.

	BRIEF	**ORB**	**BRISK**	**FREAK**	**BinBoost** ($d = 8$)	**BinBoost** ($d = 16$)	**BinBoost** ($d = 32$)
fold 1	77,19	77,03	76,41	**78,44**	77,03	76,41	76,56
fold 2	**78,51**	77,58	76,94	77,12	76,96	76,33	75,39
fold 3	75,23	75,08	71,96	**76,02**	75,55	74,92	73,83
fold 4	78,03	77,58	76,64	78,05	77,89	76,95	**78,35**
fold 5	75,85	73,53	73,53	**76,96**	76,02	75,86	76,64
Média	76,96 ± 1	76,16 ± 2	75,10 ± 1	**77,32** ± 1	76,69 ± 1	76,09 ± 1	76,15 ± 1

Tabela 5.15. Resultados de classificação (Acc % e desvio-padrão) sobre o subconjunto de validação, para a base de dados *Pornography*, com SVM $C = 1$.

	BRIEF	**ORB**	**BRISK**	**FREAK**	**BinBoost** ($d = 8$)	**BinBoost** ($d = 16$)	**BinBoost** ($d = 32$)
fold 1	84,69	84,53	82,81	81,88	**84,69**	84,22	84,53
fold 2	83,65	84,27	82,09	82,72	83,64	**84,73**	84,42
fold 3	83,95	83,34	80,68	80,69	83,49	83,64	**84,42**
fold 4	83,49	84,58	82,56	82,25	84,58	84,58	**85,35**
fold 5	83,64	84,27	82,56	80,85	84,42	**85,20**	84,73
Média	83,89 ± 0,5	84,20 ± 0,5	82,14 ± 1	81,68 ± 1	84,16 ± 0,5	84,47 ± 0,5	**84,69** ± 0,5

Tabela 5.16. Resultados de classificação (Acc % e desvio-padrão) sobre o conjunto de teste, da base de dados *Pornography*, utilizando o *framework* de combinação de classificadores [Faria et al., 2014].

	#Classificadores Combinados	Acc. (%)
fold 1	14	86,40
fold 2	11	82,50
fold 3	12	84,38
fold 4	14	83,75
fold 5	11	85,63
Média	-	**84,53** ± 1

5.3.3 Resultados do Descritor de Vídeo Proposto - BossaNova Video Descriptor

Nestes experimentos, o objetivo é avaliar o descritor de vídeo proposto, *BossaNova Video Descriptor* (VD), para o contexto de detecção de pornografia e compará-lo com métodos que também utilizaram a representação intermediária BossaNova. Para isto, os parâmetros da representação BossaNova também foram mantidos como em Avila et al. [2013]: $M = 256$, $B = 10$, $\lambda_{min} = 0{,}0$, $\lambda_{max} = 3{,}0$.

A Tabela 5.17 apresenta uma comparação com os resultados obtidos com o descritor de vídeo proposto, os resultados da abordagem base com descritores binários e resultados da literatura sobre a base de dados *Pornography*. É possível notar uma melhora considerável obtida com o BossaNova VD, alcançando 90,9% de acurácia com o descritor binário BinBoost de dimensionalidade $d = 16$. Além disso, é importante observar que este resultado supera o melhor resultado publicado da literatura, obtido utilizando descritores HueSIFT [Avila et al., 2013].

A comparação com a abordagem base proposta é particularmente relevante, pois nela são empregados os mesmos descritores binários (BRIEF, ORB, BRISK, FREAK e BinBoost) com a mesma configuração de parâmetros. É possível notar uma melhoria de (até) 2,8% quando comparada à representação intermediária BossaNova com o descritor BossaNova VD. Confirmando, assim, as vantagens introduzidas pelo descritor de vídeo proposto.

Também, é apresentada na Tabela 5.18, uma comparação do descritor de vídeo proposto com uma abordagem utilizando *pooling* global por vídeo. Ou seja, criando apenas um BossaNova por vídeo utilizando todas as características locais extraídas dos quadros-chaves. É possível observar que o descritor de vídeo proposto apresenta valores de acurácia superiores quando comparado com a abordagem usando *pooling*

Tabela 5.17. Resultados de classificação de vídeos (Acc % e desvio-padrão) do descritor de vídeo proposto, *BossaNova Video Descriptor* (VD), e os resultados publicados sobre a base de dados *Pornography*.

	Abordagem	**Acc. (%)**
Resultados da literatura	BossaNova (HueSIFT) [Avila et al., 2013]	89.5 ± 1
Abordagem base	BossaNova (BRIEF)	86,03 ± 3
	BossaNova (ORB)	86,79 ± 3
	BossaNova (BRISK)	88,65 ± 2
	BossaNova (FREAK)	86,90 ± 3
	BossaNova (BinBoost $d = 8$)	87,28 ± 2
	BossaNova (BinBoost $d = 16$)	89,40 ± 2
	BossaNova (BinBoost $d = 32$)	89,02 ± 2
Experimentos realizados	BossaNova VD (BRIEF)	89,03 ± 1
	BossaNova VD (ORB)	89,02 ± 1
	BossaNova VD (BRISK)	89,27 ± 1
	BossaNova VD (FREAK)	89,66 ± 2
	BossaNova VD (BinBoost $d = 8$)	90,77 ± 2
	BossaNova VD (BinBoost $d = 16$)	**90,90** ± 1
	BossaNova VD (BinBoost $d = 32$)	89,41 ± 2

global para todas as características locais utilizadas.

Com o objetivo de comparar diferentes funções de combinação para o descritor de vídeo proposto, foram realizados experimentos utilizando as funções: (i) Max, selecionando o valor máximo existente em $z^i_{m,b}$ e t^i_m; (ii) Min, selecionando o valor mínimo existente em $z^i_{m,b}$ e t^i_m; e (iii) Soma, somando os valores existentes em $z^i_{m,b}$ tanto quanto em t^i_m. A Tabela 5.19 apresenta os experimentos realizados. É possível observar que a função Mediana se sobressai em relação as demais funções, menos para a abordagem utilizando o descritor ORB, alcançando um valor de acurácia superior e mantendo um desvio padrão mais baixo. A função Soma apresenta um valor de acurácia ligeiramente superior para o descritor ORB, porém triplicando o desvio padrão quando comparada a função Mediana.

A Tabela 5.20 apresenta uma comparação de tempo para a criação do descritor de vídeo proposto. O cálculo é feito somando os valores de tempo médio de extração dos descritores com o tempo médio para criar a representação intermediária BossaNova (valores disponíveis na Tabela 5.12). Este valor é multiplicado pela quantidade média de tomadas por vídeo (disponível na Tabela 5.1). Por fim, é somado o tempo de combinação entre as representações intermediárias. Em média, este tempo, para um

Tabela 5.18. Resultados de classificação de vídeos (Acc % e desvio-padrão) do descritor de vídeo proposto, *BossaNova Video Descriptor* (VD), e uma abordagem utilizando *pooling* global por vídeo sobre a base de dados *Pornography*.

	Abordagem	**Acc. (%)**
Pooling global	BossaNova (BRIEF)	80,30 ± 3
	BossaNova (ORB)	78,81 ± 3
	BossaNova (BRISK)	79,93 ± 2
	BossaNova (FREAK)	79,56 ± 2
	BossaNova (BinBoost $d = 8$)	77,06 ± 1
	BossaNova (BinBoost $d = 16$)	77,31 ± 1
	BossaNova (BinBoost $d = 32$)	77,93 ± 1
BossaNova Video Descriptor	BossaNova VD (BRIEF)	89,03 ± 1
	BossaNova VD (ORB)	89,02 ± 1
	BossaNova VD (BRISK)	89,27 ± 1
	BossaNova VD (FREAK)	89,66 ± 2
	BossaNova VD (BinBoost $d = 8$)	90,77 ± 2
	BossaNova VD (BinBoost $d = 16$)	**90,90** ± 1
	BossaNova VD (BinBoost $d = 32$)	89,41 ± 2

Tabela 5.19. Resultados de classificação de vídeos (Acc % e desvio-padrão) do descritor de vídeo proposto, *BossaNova Video Descriptor* (VD), utilizando diversas funções de combinação sobre a base de dados *Pornography*.

Abordagens	**Max**	**Min**	**Soma**	**Mediana**
BossaNova VD (BRIEF)	88,16 ± 3	80,92 ± 4	88,66 ± 2	**89,03** ± 1
BossaNova VD (ORB)	89,28 ± 3	84,42 ± 3	**89,28** ± 3	89,02 ± 1
BossaNova VD (BRISK)	87,16 ± 2	82,30 ± 4	88,66 ± 1	**89,27** ± 1
BossaNova VD (FREAK)	87,66 ± 2	85,66 ± 1	88,04 ± 1	**89,66** ± 2
BossaNova VD (BinBoost $d = 8$)	89,28 ± 3	82,05 ± 3	87,66 ± 1	**90,77** ± 2
BossaNova VD (BinBoost $d = 16$)	87,79 ± 2	82,42 ± 3	88,54 ± 2	**90,90** ± 1
BossaNova VD (BinBoost $d = 32$)	88,28 ± 2	81,92 ± 2	88,15 ± 1	**89,41** ± 2

dicionário visual de tamanho $M = 256$, é 0,012 milissegundos.

Além disso, foi realizada uma análise para verificar as falhas de classificação do descritor BossaNova VD. Os vídeos não-pornográficos classificados erroneamente correspondem aos vídeos relacionados ao sub-grupo *"difícil"*, como vídeos de amamentação, vídeos de pessoas dando banho em bebês e cenas de praia (que envolvem alta exposição de pele). Além disso, a análise dos vídeos pornográficos classificados incorretamente revelou que o descritor de vídeo proposto apresenta dificuldades com vídeos de baixa qualidade ou vídeos cujo conteúdo pornográfico não é explícito. Esta

Tabela 5.20. Comparação de tempo (em segundos) para criação do descritor de vídeo proposto. O cálculo é feito somando os valores de tempo médio de extração dos descritores com o tempo médio para criar a representação intermediária BossaNova. Este valor é multiplicado pela quantidade média de tomadas por vídeo (20,6) e, por fim, é somado o tempo de combinação entre as representações intermediárias (0,012).

Abordagem	**Tempo Total**
BossaNova VD (BRIEF)	15,26
BossaNova VD (ORB)	15,84
BossaNova VD (BRISK)	30,29
BossaNova VD (FREAK)	23,70
BossaNova VD (BinBoost $d = 8$)	16,70
BossaNova VD (BinBoost $d = 16$)	21,44
BossaNova VD (BinBoost $d = 32$)	31,53

mesma dificuldade também é relatada por Avila et al. [2013].

Variação do Tamanho do Dicionário Visual

O impacto do tamanho M do dicionário visual no desempenho da classificação do descritor BossaNova VD + BinBoost ($d = 16$) é mostrado na Tabela 5.21, mostrando uma melhor acurácia para dicionários visuais maiores.

Tabela 5.21. Resultados de classificação de vídeos (Acc % e desvio-padrão) do descritor de vídeo proposto, *BossaNova Video Descriptor* (VD) + BinBoost ($d = 16$), relacionados à variação do tamanho M do dicionário visual na base de dados *Pornography*.

Dicionário	**Acc. (%)**
$M = 64$	87,66 ± 1
$M = 128$	89,16 ± 2
$M = 256$	90,90 ± 1
$M = 512$	91,28 ± 1
$M = 1024$	**92,02** ± 1
$M = 2048$	91,28 ± 2
$M = 4096$	91,65 ± 2

Como apresentado na Tabela 5.17, os resultados alcançados correspondem aos parâmetros utilizados de [Avila et al., 2013], desta maneira estes parâmetros não foram ajustados para o descritor proposto. Portanto, o descritor BossaNova VD pode alcançar um resultado ainda maior (92,02%), com um dicionário visual de tamanho $M = 1024$,

como pode ser visto na Tabela 5.21.

5.4 Considerações

Durante a elaboração deste trabalho, diversos experimentos foram realizados. Os experimentos relacionados à base de dados PASCAL VOC 2007 foram importantes para estudar o comportamento dos descritores binários em tarefas de classificação. A partir destes experimentos, foi decidido qual descritor binário teria um estudo mais aprofundado de seus parâmetros (BinBoost). Os experimentos relacionados ao contexto de detecção de pornografia, usando a abordagem de classificação base com descritores binários, foram importantes para estudar o comportamento da metodologia proposta e principalmente para validar, novamente, o uso do descritor binário BinBoost. Além disso, os experimentos relacionados ao *framework* de combinação de classificadores foram importantes para mostrar que uma classificação tardia (em relação ao classificador) não traria ganhos para o problema estudado com a abordagem proposta. A partir disso, uma solução de combinação de representações intermediárias foi proposta e então novos experimentos foram realizados, dando origem ao descritor de vídeo BossaNova VD. Os resultados obtidos validaram o descritor proposto para o contexto de detecção de pornografia, além de validar o uso dos descritores binários para tarefas de reconhecimento visual.

Capítulo 6

Conclusões e Trabalhos Futuros

Com o crescimento exponencial de conteúdos inapropriados na Internet, como pornografia, surge uma necessidade de se detectar e filtrar tal tipo de material. O motivo disto é dado pelo fato de que esse tipo de conteúdo é frequentemente proibido em certos ambientes (por exemplo, escolas e locais de trabalho) ou para certos públicos (crianças).

Em vista disso, neste trabalho, foi apresentada uma metodologia para detecção de vídeos pornográficos. A metodologia proposta integrou as vantagens dos conceitos apresentados nos trabalhos de referência na área de detecção de pornografia com recentes descritores de características locais de imagens, contribuindo para o aprimoramento do estado da arte desta área de aplicação.

Mais especificamente, este trabalho se concentrou na classificação supervisionada de vídeos, pornográficos e não-pornográficos, baseada em características locais de imagens codificadas por representações intermediárias. A descrição das características locais foi feita utilizando descritores binários, uma alternativa de baixa complexidade, e para a codificação destas características foi utilizada a recente representação intermediária BossaNova, uma extensão do modelo *Bag-of-Words* que preserva de uma maneira mais rica a informação visual.

Foram apresentadas duas abordagens para detecção de pornografia. A primeira, tratada neste trabalho como abordagem de classificação base, é baseada no trabalho de Avila et al. [2013] se diferenciando em duas etapas: (i) extração de características locais, onde faz o uso de descritores binários; e (ii) criação do dicionário visual, utilizando o algoritmo de clusterização *k-medians* e distância de Hamming como medida de distância. Os resultados obtidos com essa abordagem base demonstraram estar bem próximos ao melhor resultado relatado pela literatura. Além disso, foram realizados experimentos com um *framework* de combinação de classificadores para combinar os

diferentes resultados alcançados por cada descritor binário utilizado.

A segunda abordagem se baseia em uma combinação de representações intermediárias, dando origem ao descritor de vídeo BossaNova VD. Intuitivamente, este novo descritor de vídeo representa a distância média para cada característica local em relação às palavras visuais do dicionário visual. Os resultados obtidos validaram o descritor proposto para o contexto de detecção de pornografia apresentando resultados com qualidade superior em relação às demais abordagens encontradas na literatura.

Além disso, foram realizados experimentos sobre a base de dados PASCAL VOC 2007 com o objetivo de validar o uso de descritores binários para tarefas de reconhecimento visual. A solução proposta apresentou resultados com qualidade superior em relação às demais abordagens encontradas na literatura que também empregaram descritores binários.

Com relação aos trabalhos futuros, o presente trabalho pode ser continuado explorando-se os seguintes aspectos:

- Estudo sobre a variação dos parâmetros dos descritores binários.
- Estudo sobre a variação de parâmetros da representação intermediária BossaNova.
- Verificar a adequação da abordagem base proposta, assim como do descritor de vídeo proposto, a outros gêneros de vídeos, tais como desenho animado, esportivo, filme.
- Verificar a adequação da abordagem base proposta, assim como do descritor de vídeo proposto, a outros contextos de reconhecimento visual, tais como detecção de ações e atividades.
- Investigar o uso de outras técnicas de classificação, como por exemplo o algoritmo AdaBoost [Freund & Schapire, 1997], que utiliza a ideia de ponderar a saída de um classificador com o objetivo de alimentar o próximo classificador.
- Combinação entre descritores binários e descritores não-binários, como HueSIFT.
- Aplicação de técnicas de *Deep learning* [Arel et al., 2010] para uma melhor classificação dos dados.

6.1 Publicações

Conferências Internacionais

- Caetano, C.; Avila, S.; Guimarães, S. & de A. Araújo, A.. Representing Local Binary Descriptors with BossaNova for Visual Recognition. Em Proceedings of the 29th Annual ACM Symposium on Applied Computing (SAC'14), pp. 49–54, Gyeongju, Coreia do Sul, 2014.

- Caetano, C.; Avila, S.; Guimarães, S. & de A. Araújo, A.. Pornography Detection Using A New Video Descriptor. Em 22nd European Signal Processing Conference (EUSIPCO), pp. 1681–1685, Lisboa, Portugal, 2014.

Referências Bibliográficas

Agarwal, S.; Awan, A. & Roth, D. (2004). Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(11):1475--1490. 2

Agrawal, M.; Konolige, K. & Blas, M. (2008). Censure: Center surround extremas for realtime feature detection and matching. Em Forsyth, D.; Torr, P. & Zisserman, A., editores, *Computer Vision (ECCV) 2008*, volume 5305 of *Lecture Notes in Computer Science*, pp. 102–115. Springer Berlin Heidelberg. 10

Aizerman, M. A.; Braverman, E. A. & Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. Em *Automation and Remote Control,*, number 25 in Automation and Remote Control„ pp. 821–837. 24

Arandjelovic, R. & Zisserman, A. (2012). Three things everyone should know to improve object retrieval. Em *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2911–2918. 12

Arel, I.; Rose, D. & Karnowski, T. (2010). Deep machine learning - a new frontier in artificial intelligence research [research frontier]. *Computational Intelligence Magazine, IEEE*, 5(4):13–18. 64

Avila, S. (2013). *Extended Bag-Of-Words Formalism For Image Classification*. Tese de doutorado, Federal University of Minas Gerais. 19, 32

Avila, S.; Thome, N.; Cord, M.; Valle, E. & Araújo, A. (2013). Pooling in image representation: the visual codeword point of view. *CVIU*, 117(5):453--465. xvii, 2, 3, 18, 20, 21, 23, 24, 31, 32, 33, 35, 36, 37, 39, 41, 43, 44, 45, 46, 48, 52, 53, 54, 58, 59, 61, 63

Avila, S.; Thome, N.; Cord, M.; Valle, E. & de A. Araújo, A. (2011). BOSSA: Extended bow formalism for image classification. Em *ICIP*, pp. 2909--2912. 21, 23, 24, 41, 43, 45, 48

Avila, S.; Thome, N.; Cord, M.; Valle, E. & de A. Araújo, A. (2012). Bossanova at imageclef 2012 flickr photo annotation task. Em *Working Notes of the CLEF*. 7, 22, 23

Bay, H.; Ess, A.; Tuytelaars, T. & Van Gool, L. (2008). Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346--359. 7, 10, 11, 13

Beaudet, P. R. (1978). Rotationally invariant image operators. Em *Proceedings of the 4th International Joint Conference on Pattern Recognition*, pp. 579--583, Kyoto, Japan. 10

Bkassiny, M.; Li, Y. & Jayaweera, S. (2013). A survey on machine-learning techniques in cognitive radios. *Communications Surveys Tutorials, IEEE*, 15(3):1136–1159. 24

Boureau, Y.-L.; Bach, F.; LeCun, Y. & Ponce, J. (2010). Learning mid-level features for recognition. Em *CVPR*, pp. 2559--2566. 19

Bradley, P. S.; Mangasarian, O. L. & Street, W. N. (1997). Clustering via concave minimization. Em *Advances in Neural Information Processing Systems -9*, pp. 368--374. MIT Press. 37

Bradski, G. (2000). The opencv library. *Dr. Dobb's Journal of Software Tools*. 47

Caetano, C.; Avila, S.; Guimarães, S. & de A. Araújo, A. (2014). Representing local binary descriptors with BossaNova for visual recognition. Em *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, SAC '14, pp. 49--54. 24, 38

Calonder, M.; Lepetit, V.; Strecha, C. & Fua, P. (2010). BRIEF: binary robust independent elementary features. Em *ECCV*, pp. 778--792. 14, 47

Canclini, A.; Cesana, M.; A., R.; Tagliasacchi, M.; Ascenso, J. & R., C. (2013). Evaluation of low-complexity visual feature detectors and descriptors. Em *Int. Conf. on Digital Signal Processing*. 8, 9, 10, 11

Chatfield, K.; Lemtexpitsky, V.; Vedaldi, A. & Zisserman, A. (2011). The devil is in the details: an evaluation of recent feature encoding methods. Em *BMVC*. 7, 19, 20, 50

Chatzilari, E.; Liaros, G.; Nikolopoulos, S. & Kompatsiaris, Y. (2013). A comparative study on mobile visual recognition. Em *MLDM*, volume 7988, pp. 442--457. Springer Berlin Heidelberg. 49, 50

Cristianini, N. & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA. 24

Csurka, G.; Dance, C. R.; Fan, L.; Willamowski, J. & Bray, C. (2004). Visual categorization with bags of keypoints. Em *In Workshop on Statistical Learning in Computer Vision, ECCV*, pp. 1--22. 7

Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection. Em *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pp. 886–893 vol. 1. 11

Deselaers, T.; Pimenidis, L. & Ney, H. (2008). Bag-of-visual-words models for adult image classification and filtering. Em *ICPR*. 2, 24, 30, 34, 37

Duda, R. O. & Hart, P. E. (2000). *Pattern Classification and Scene Analysis*. John Wiley & Sons Inc, 2 edição. 24

Endeshaw, T.; Garcia, J. & Jakobsson, A. (2008). Classification of indecent videos by low complexity repetitive motion detection. Em *Applied Imagery Pattern Recognition Workshop, 2008. AIPR '08. 37th IEEE*, pp. 1–7. 32

Everingham, M.; Van Gool, L.; Williams, C.; Winn, J. & Zisserman, A. (2007). The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. 41, 42, 45

Faria, F. A.; Dos Santos, J. A.; Rocha, A. & Torres, R. D. S. (2014). A framework for selection and fusion of pattern classifiers in multimedia recognition. *Pattern Recogn. Lett.*, 39:52--64. xvii, xviii, 54, 55, 56, 58

Forsyth, D. & Fleck, M. (1996). Identifying nude pictures. Em *Applications of Computer Vision, 1996. WACV '96., Proceedings 3rd IEEE Workshop on*, pp. 103–108. 2, 27

Forsyth, D. A. & Fleck, M. M. (1997). Body plans. Em *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 678--683. 2, 27

Forsyth, D. A. & Fleck, M. M. (1999). Automatic detection of human nudes. *Int. J. Comput. Vision*, 32(1):63--77. 2, 27

Freund, Y. & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119--139. 17, 29, 64

Gaidon, A.; Harchaoui, Z. & Schmid, C. (2012). Recognizing activities with cluster-trees of tracklets. Em Bowden, R.; Collomosse, J. P. & Mikolajczyk, K., editores, *BMVC 2012 - British Machine Vision Conference*, pp. 30.1–30.13, Guildford, United Kingdom. BMVA Press. 24

Gauglitz, S.; Höllerer, T. & Turk, M. (2011). Evaluation of interest point detectors and feature descriptors for visual tracking. *IJCV*, 94(3):335--360. 9

Gemert, J. C.; Geusebroek, J.-M.; Veenman, C. J. & Smeulders, A. W. (2008). Kernel codebooks for scene categorization. Em *Proceedings of the 10th European Conference on Computer Vision: Part III*, ECCV '08, pp. 696--709, Berlin, Heidelberg. Springer-Verlag. 23

Ghahramani, Z. (2004). Unsupervised learning. Em Bousquet, O.; Luxburg, U. & Rätsch, G., editores, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pp. 72–112. Springer Berlin Heidelberg. 23, 24

Gosselin, P. H.; Cord, M. & Philipp-Foliguet, S. (2008). Combining visual dictionary, kernel-based similarity and learning strategy for image category retrieval. *Computer Vision and Image Understanding*, 110(3):403 – 417. Similarity Matching in Computer Vision and Multimedia. 7

Harris, C. & Stephens, M. (1988). A combined corner and edge detector. Em *In Proc. of Fourth Alvey Vision Conference*, pp. 147--151. 9

Hu, W.; Wu, O.; Chen, Z.; Fu, Z. & Maybank, S. (2007). Recognition of pornographic web pages by classifying texts and images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):1019--1034. 2, 29

Hu, W.; Zuo, H.; Wu, O.; Chen, Y.; Zhang, Z. & Suter, D. (2011). Recognition of adult images, videos, and web page bags. *ACM Trans. Multimedia Comput. Commun. Appl.*, 7S(1):28:1--28:24. 29

Jain, A. K. & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. 37

Jain, R. K. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley, 1 edição. 46

Jansohn, C.; Ulges, A. & Breuel, T. M. (2009). Detecting pornographic video content by combining image features with motion information. Em *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, pp. 601--604, New York, NY, USA. ACM. 32

Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C. & Black, M. (2013). Towards understanding action recognition. Em *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 3192–3199. 24

Jones, M. J. & Rehg, J. M. (2002). Statistical color models with application to skin detection. *Int. J. Comput. Vision*, 46(1):81--96. 2, 28, 31, 34

Jurie, F. & Triggs, B. (2005). Creating efficient codebooks for visual recognition. Em *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pp. 604–610 Vol. 1. 11

Ke, Y. & Sukthankar, R. (2004). Pca-sift: a more distinctive representation for local image descriptors. Em *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pp. II–506–II–513 Vol.2. 11, 12

Kim, J. & Grauman, K. (2011). Boundary preserving dense local regions. Em *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1553–1560. 11

Kläser, A.; Marszałek, M. & Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. Em *British Machine Vision Conference*, pp. 995--1004. 24

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Em *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, pp. 3--24, Amsterdam, The Netherlands, The Netherlands. IOS Press. 24

Krapac, J.; Verbeek, J. & Jurie, F. (2011). Modeling spatial layout with fisher vectors for image categorization. Em *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 1487–1494. 23

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience. 54

Laptev, I. (2005). On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107--123. 33

Laptev, I.; Marszalek, M.; Schmid, C. & Rozenfeld, B. (2008). Learning realistic human actions from movies. Em *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8. 24

Lazebnik, S.; Schmid, C. & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. Em *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pp. 2169–2178. 7, 19, 23

Lee, J.-S.; Kuo, Y.-M.; Chung, P.-C. & Chen, E.-L. (2007). Naked image detection based on adaptive and extensible skin color model. *Pattern Recogn.*, 40(8):2261--2270. 28, 29

Lee, J.-S.; Yu, F.-S. & Huang, K.-Y. (2013). Pornography detection based on morphological features. *International Journal of Computer, Consumer and Control (IJ3C)*, 2:56--64. 29, 37

Leutenegger, S.; Chli, M. & Siegwart, R. (2011). BRISK: Binary robust invariant scalable keypoints. Em *ICCV*, pp. 2548--2555. 16, 47

Lindeberg, T. (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30:79--116. 9

Liu, Y.; Wang, X.; Zhang, Y. & Tang, S. (2011). Fusing audio-words with visual features for pornographic video detection. Em *Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on*, pp. 1488–1493. 34

Lloyd, S. (2006). Least squares quantization in pcm. *IEEE Trans. Inf. Theor.*, 28(2):129--137. 19

Lopes, A.; Avila, S.; Peixoto, A.; Oliveira, R.; Coelho, M. & de A. Araújo, A. (2009a). Nude detection in video using bag-of-visual-features. Em *SIBGRAPI*. 2, 24, 31

Lopes, A.; Avila, S.; Peixoto, A.; Oliveira, R. & de A. Araújo, A. (2009b). A bag-of-features approach based on Hue-SIFT descriptor for nude detection. Em *EUSIPCO*. 1, 2, 24, 30

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91--110. 7, 10, 11, 12

Mair, E.; Hager, G. D.; Burschka, D.; Suppa, M. & Hirzinger, G. (2010). Adaptive and generic corner detection based on the accelerated segment test. Em *Proceedings of the 11th European Conference on Computer Vision: Part II*, ECCV'10, pp. 183--196, Berlin, Heidelberg. Springer-Verlag. 9

Mikolajczyk, K. & Schmid, C. (2001). Indexing based on scale invariant interest points. Em *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pp. 525–531 vol.1. 9

Mikolajczyk, K. & Schmid, C. (2004). Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63--86. 9, 10

Mikolajczyk, K. & Schmid, C. (2005). A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630. 9, 11, 12

Miksik, O. & Mikolajczyk, K. (2012). Evaluation of local detectors and descriptors for fast feature matching. Em *ICPR*, pp. 2681--2684. 9

Moravec, H. P. (1980). *Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover*. Tese de doutorado, Stanford, Stanford, CA, USA. AAI8024717. 9

Morel, J. & Yu, G. (2011). Is sift scale invariant? *Inverse Problems and Imaging*, 5(1):115--136. 12

Ortiz, R. (2012). FREAK: Fast retina keypoint. Em *CVPR*, pp. 510--517. 17, 47

Perronnin, F.; Sánchez, J. & Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. Em *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pp. 143--156, Berlin, Heidelberg. Springer-Verlag. 23

Picard, D.; Thome, N. & Cord, M. (2013). Jkernelmachines: A simple framework for kernel machines. *Journal of Machine Learning Research*, 14:1417–1421. 48

Rea, N.; Lacey, G.; Lambe, C. & Dahyot, R. (2006). Multimodal periodicity analysis for illicit content detection in videos. Em *Visual Media Production, 2006. CVMP 2006. 3rd European Conference on*, pp. 106–114. 33

Ries, C. & Lienhart, R. (2014). A survey on visual adult image recognition. *Multimedia Tools and Applications*, 69(3):661–688. 2, 27, 30, 31, 37

Rokach, L. (2010). Ensemble-based classifiers. *Artif. Intell. Rev.*, 33(1-2):1--39. 55

Rosin, P. (1999). Measuring corner properties. Em *Computer Vision & Image Understanding, Vol.73, No.2*, pp. 291--307. 15

Rosten, E. & Drummond, T. (2006). Machine learning for high-speed corner detection. Em *Proceedings of the 9th European Conference on Computer Vision - Volume Part I*, ECCV'06, pp. 430--443, Berlin, Heidelberg. Springer-Verlag. 9

Rowley, H. A.; Jing, Y. & Baluja, S. (2006). Large scale image-based adult-content filtering. Em Ranchordas, A.; Araújo, H. & Encarnação, B., editores, *International Conference on Computer Vision Theory and Applications (VISAPP)*, pp. 290–296. INSTICC - Institute for Systems and Technologies of Information, Control and Communication. 2, 28

Rublee, E.; Rabaud, V.; Konolige, K. & Bradski, G. R. (2011). ORB: An efficient alternative to SIFT or SURF. Em *ICCV*, pp. 2564--2571. 15, 47

Sahidullah, M. & Saha, G. (2012). Design, analysis and experimental evaluation of block based transformation in {MFCC} computation for speaker recognition. *Speech Communication*, 54(4):543 – 565. 33

SáNchez, J.; Perronnin, F. & De Campos, T. (2012). Modeling the spatial layout of images beyond spatial pyramids. *Pattern Recogn. Lett.*, 33(16):2216--2223. 23

Schmid, C.; Mohr, R. & Bauckhage, C. (2000). Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172. 9

Scholkopf, B. & Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA. 24

Schuldt, C.; Laptev, I. & Caputo, B. (2004). Recognizing human actions: a local svm approach. Em *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pp. 32–36 Vol.3. 24

Short, M. B.; Black, L.; Smith, A. H.; Wetterneck, C. T. & Wells, D. E. (2012). A review of internet pornography use research: Methodology and content from the past 10 years. *Cyberpsy., Behavior, and Soc. Networking*, 15(1):13–23. 2

Sivic, J. & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. Em *ICCV*, pp. 1470--1477. 7, 18

Souza, F.; Valle, E.; Cámara-Chávez, G. & Araújo, A. d. A. (2012). An evaluation on color invariant based local spatiotemporal features for action recognition. Em *25th Conference on Graphics, Patterns and Images*. 33

Steel, C. (2012). The mask-sift cascading classifier for pornography detection. Em *Internet Security (WorldCIS), 2012 World Congress on*, pp. 139–142. 2, 31

Trajkovic, M. & Hedley, M. (1998). Fast corner detection. *Image and Vision Computing*, 16(2):75--87. 9

Truong, B. T. & Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1). 35

Trzcinski, T.; Christoudias, M.; Fua, P. & Lepetit, V. (2013). Boosting binary keypoint descriptors. Em *CVPR*, pp. 2874–2881. 17, 47

Tuytelaars, T. (2010). Dense interest points. Em *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2281–2288. 8, 11

Tuytelaars, T. & Mikolajczyk, K. (2008). Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177--280. 8, 9, 10

Ulges, A.; Schulze, C.; Borth, D. & Stahl, A. (2012). Pornography detection in video benefits (a lot) from a multi-modal approach. Em *Proceedings of the 2012 ACM International Workshop on Audio and Multimedia Methods for Large-scale Video Analysis*, AMVA '12, pp. 21--26, New York, NY, USA. ACM. 34

Ulges, A. & Stahl, A. (2011). Automatic detection of child pornography using color visual words. Em *ICME*. 2, 24, 31

Valle, E.; de Avila, S. E. F.; da Luz Jr., A.; de Souza, F. D. M.; de M. Coelho, M. & de Albuquerque Araújo, A. (2011). Content-based filtering for video sharing social networks. *CoRR*, abs/1101.2427. 2

Valle, E.; de Avila, S. E. F.; da Luz Jr., A.; de Souza, F. D. M.; de M. Coelho, M. & de Albuquerque Araújo, A. (2012). Content-based filtering for video sharing social networks. Em *Anais do SBSeg 2012 - XII Simpósio em Segurança da Informação e de Sistemas Computacionais : WFC - Workshop de Forense Computacional*, pp. 625–638, Curitiba, PR. 33

Van de Sande, K. E. A.; Gevers, T. & Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582–1596. 3, 12

Vapnik, V. & Lerner, A. (1963). Pattern Recognition using Generalized Portrait Method. *Automation and Remote Control*, 24. 24

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA. 24

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience, Ney York. 24

Wang, H.; Ullah, M. M.; Kläser, A.; Laptev, I. & Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. Em *University of Central Florida, U.S.A.*. 11, 24

Wu, O.; Zuo, H.; Hu, W.; Zhu, M. & Li, S. (2008). Recognizing and filtering web images based on people's existence. Em *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on*, volume 1, pp. 648–654. 29

Yang, J.; Jiang, Y.-G.; Hauptmann, A. G. & Ngo, C.-W. (2007). Evaluating bag-of-visual-words representations in scene classification. Em *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, MIR '07, pp. 197--206, New York, NY, USA. ACM. 2

Yang, J.; Yu, K.; Gong, Y. & Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. Em *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1794–1801. 23

Yu, J.-J. & Han, S.-W. (2014). Skin detection for adult image identification. Em *Advanced Communication Technology (ICACT), 2014 16th International Conference on*, pp. 645–648. 2, 29

Özgür, A. (2004). Supervised and unsupervised machine learning techniques for text document categorization. Relatório técnico, Boğaziçi University. 24

Zhang, Y.; Zhu, C.; Bres, S. & Chen, L. (2013). Encoding local binary descriptors by bag-of-features with hamming distance for visual object categorization. Em *ECIR*, pp. 630--641. 49, 50

Zheng, H.; Daoudi, M. & Jedynak, B. (2004). Blocking Adult Images Based on Statistical Skin Detection. *Electronic Letters on Computer Vision and Image Analysis*, 4(2):1--14. 2, 28

Zhou, X.; Yu, K.; Zhang, T. & Huang, T. S. (2010). Image classification using super-vector coding of local image descriptors. Em *Proceedings of the 11th European Conference on Computer Vision: Part V*, ECCV'10, pp. 141--154, Berlin, Heidelberg. Springer-Verlag. 23, 37

Zhu, X. (2006). Semi-supervised learning literature survey. 24

Zuo, H.; Hu, W. & Wu, O. (2010). Patch-based skin color detection and its application to pornography image filtering. Em *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 1227--1228, New York, NY, USA. ACM. 29

Zuo, H.; Wu, O.; Hu, W. & Xu, B. (2008). Recognition of blue movies by fusion of audio and video. Em *Multimedia and Expo, 2008 IEEE International Conference on*, pp. 37–40. 33