

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

DETECÇÃO DE VÍDEOS NÃO COLABORATIVOS COM BASE NO
CONTEÚDO VISUAL EM REDES SOCIAIS PARA
COMPARTILHAMENTO DE VÍDEO

Antonio da Luz Júnior

Belo Horizonte
Agosto de 2012.

**DETECÇÃO DE VÍDEOS NÃO COLABORATIVOS COM BASE
NO CONTEÚDO VISUAL EM REDES SOCIAIS PARA
COMPARTILHAMENTO DE VÍDEO**

ANTONIO DA LUZ JÚNIOR

**DETECÇÃO DE VÍDEOS NÃO COLABORATIVOS COM BASE NO
CONTEÚDO VISUAL EM REDES SOCIAIS PARA
COMPARTILHAMENTO DE VÍDEO**

Tese 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 Doutor em Ciência da Computação.

ORIENTADOR: ARNALDO DE ALBUQUERQUE ARAÚJO

Co-ORIENTADOR: EDUARDO VALLE

Belo Horizonte

Agosto de 2012

Luz Júnior, Antônio da.

L979d Detecção de vídeos não colaborativos com base no conteúdo visual em redes sociais para compartilhamento de vídeo / Antônio da Luz Júnior. — Belo Horizonte, 2012.
xxvi, 70 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientador: Arnaldo de Albuquerque Araújo.
Coorientador: Eduardo Valle.

1. Computação - Teses. 2. Redes sociais on-line – Teses. 3. Semântica – Teses. 4. Vídeo digital. I. Orientador. II. Coorientador. III. Título

519.6*04(043)



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

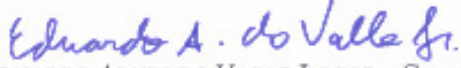
FOLHA DE APROVAÇÃO

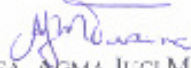
Detecção de vídeos não-colaborativos com base no conteúdo visual em redes
sociais para compartilhamento de vídeo

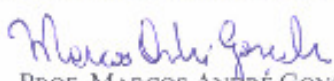
ANTONIO DA LUZ JUNIOR

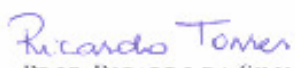
Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:


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


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


PROF. AGMA JUCI MACHADO TRAINA
Instituto de Ciências Matemáticas e Computação - USP


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


PROF. RICARDO DA SILVA TORRES
Instituto de Computação - UNICAMP


PROF. WAGNER MEIRA JÚNIOR
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 20 de abril de 2012.

Agradecimentos

Agradeço a Deus e a Nossa Senhora Aparecida pelas bênçãos que sempre têm me proporcionado.

A meus pais pelo apoio, confiança e amor sempre incondicionais.

Aos meus familiares e amigos pelo incentivo que sempre me deram.

Ao Prof. Arnaldo de A. Araújo pela oportunidade e confiança sem as quais nada disso teria sido possível. A sua valiosa colaboração, como um mestre experiente e sempre vigilante, nos deu a tranquilidade necessária para desenvolvermos esse trabalho, certos de que sempre tínhamos um porto-seguro ao qual recorrer.

Ao Prof. Eduardo Valle, grande mestre, por toda paciência, dedicação e orientação. Sem dúvidas teve papel fundamental no desenvolvimento deste trabalho, bem como na minha formação pessoal e profissional.

Aos Profs. Agma Juci Machado Traina, Ricardo da Silva Torres, Marcos André Gonçalves e Wagner Meira Jr., por acompanharem o desenvolvimento deste trabalho, avaliando e orientando como verdadeiros mestres.

Aos grandes amigos e incentivadores Daniel Abdalla, Eduardo Leal e Fabiano Fagundes que participaram das etapas iniciais de minha formação acadêmica, sempre me incentivaram e acreditaram que seria possível.

Aos amigos e companheiros membros do NPDI por me auxiliarem no decorrer dessa jornada, com palavras de incentivo, discussões sobre rumos de projeto, compartilhamento de material/algoritmos, acolhimento, amizade e confiança.

A Renata, Sheila e todos os demais servidores da secretaria do PPGCC por sempre me atenderem com atenção, dedicação e presteza nas inúmeras vezes que necessitei.

Ao CNPq, CAPES, FAPEMIG e FAPESP, agências de financiamento que proporcionaram o suporte financeiro necessário para o desenvolvimento do trabalho.

Resumo

A filtragem de conteúdo em redes sociais é uma tarefa em crescente demanda, devido, dentre outras razões, ao aumento da popularidade dessas redes, que estimula o aumento dos abusos, irritando usuários e prejudicando os serviços oferecidos. Neste trabalho estamos focados na detecção de vídeos não colaborativos em redes sociais para compartilhamento de vídeo. Especificamente, investigamos o quanto a análise baseada em conteúdo visual pode ajudar a detectar vídeos dos tipos *stuffing* e *spam* em *threads* de vídeo-resposta. Esta é uma tarefa bastante desafiadora, devido ao alto-nível dos conceitos semânticos envolvidos, à natureza variada das redes sociais e, um dos pontos mais importantes, à dependência ao contexto na caracterização dos vídeos não colaborativos.

Propomos uma representação sensível ao contexto, que amplia consideravelmente a qualidade da detecção em comparação aos métodos tradicionais de *bag-of-visual-features*. A metodologia é avaliada em duas bases de dados desafiadoras e os resultados demonstram a sua viabilidade.

Keywords: Filtragem de Conteúdo. *Bags of Visual Features*. Análise de Semântica Latente. Redes Sociais para Compartilhamento de Vídeo.

Abstract

In this work we are concerned with detecting non-collaborative videos in video sharing social networks. Specifically, we investigate how much visual content-based analysis can aid in detecting ballot stuffing and spam videos in threads of video responses. That is a very challenging task, because of the high-level semantic concepts involved; of the assorted nature of social networks, preventing the use of constrained a priori information; and, which is paramount, of the context-dependent nature of non-collaborative videos. Content filtering for social networks is an increasingly demanded task: due to their popularity, the number of abuses also tends to increase, annoying the user and disrupting their services.

We propose a context-aware description, which improves detection considerably in comparison with the baseline bags-of-visual-words model, by allowing us to incorporate the context of the video into the representation. Our model is evaluated in two challenging video dataset and show the feasibility of the proposed approaches.

Keywords: Content filtering. Bags of Visual Features. Latent Semantic Analysis. Video sharing social networks.

Lista de Figuras

Figura 1.1. Quadros extraídos de vídeos de uma <i>thread</i> de vídeos-resposta exemplificando a ocorrência de elementos da categoria <i>Stuffing</i> . Os vídeos que compartilham conteúdo estão identificados com uma letra entre parênteses.	7
Figura 1.2. Quadros extraídos de vídeos de uma <i>thread</i> de vídeos-resposta exemplificando a ocorrência de elementos da categoria <i>Spam</i>	8
Figura 3.1. Estrutura genérica da abordagem de <i>bag-of-visual-features</i>	38
Figura 3.2. Processo de identificação de hiperplano para separação de conjunto de dados. O hiperplano <i>h2</i> representa a melhor opção, por obter a maior distância entre o hiperplano e os vetores de suporte.....	40
Figura 4.1. Arquitetura da “solução modelo” com combinação de evidências. As etapas de treinamento e classificação possuem execução similar, sendo que, na etapa de classificação é adicionada a etapa de combinação de evidências.....	48
Figura 6.1. Arquitetura da metodologia proposta, com características de ser sensível ao contexto, utilizando a localização contextual dos vídeos, e utilizar o espaço de semântica latente para projeção das características visuais.	74
Figura 6.2. Esquema visual contrastando o vetor de características livre de contexto, baseado no <i>bag-of-visual-features</i> (a), e o vetor de características sensível ao contexto, baseado no <i>bag-of-topic-differences</i> (b).	77
Figura 7.1. Curva ROC para avaliação dos experimentos <i>Bag-of-visual-features</i> (BoVF), <i>Bag-of-visual-differences</i> (BoVD), <i>Bag-of-topics</i> (BoT) e <i>Bag-of-topics-differences</i> (BoTD), na base <i>controlada</i> , utilizando o descritor <i>SIFT</i>	95
Figura 7.2. Curva ROC para avaliação dos experimentos <i>Bag-of-visual-features</i> (BoVF), <i>Bag-of-visual-differences</i> (BoVD), <i>Bag-of-topics</i> (BoT) e <i>Bag-of-topics-differences</i> (BoTD), na base <i>controlada</i> , utilizando o descritor <i>STIP</i>	96
Figura 7.3. Curva ROC para avaliação dos experimentos <i>Bag-of-visual-features</i> (BoVF), <i>Bag-of-visual-differences</i> (BoVD), <i>Bag-of-topics</i> (BoT) e <i>Bag-of-topics-differences</i> (BoTD), na base <i>selvagem</i> , utilizando o descritor <i>SIFT</i>	99
Figura 7.4. Curva ROC para avaliação dos experimentos <i>Bag-of-visual-features</i> (BoVF), <i>Bag-of-visual-differences</i> (BoVD), <i>Bag-of-topics</i> (BoT) e <i>Bag-of-topics-differences</i> (BoTD), na base <i>selvagem</i> , utilizando o descritor <i>STIP</i>	100

Lista de Tabelas

Tabela 2.1. Relação de Redes Sociais para compartilhamento de vídeo mais populares ¹ .	16
Tabela 5.1. Quantidade de vídeos coletados por classe.	58
Tabela 5.2. Distribuição de vídeos e quantitativo de quadros por classe, utilizados nos experimentos deste trabalho.....	59
Tabela 5.3. Fórmulas para o cálculo de Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC).....	61
Tabela 5.4. Valores médios para Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC). Os melhores resultados individuais aparecem destacados em negrito.	63
Tabela 5.5. Matriz com resultados do teste-t para os valores de Coeficiente de Correlação de Matthews (MCC) com significância de 95%.	64
Tabela 5.6. Valores médios para Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC). Os melhores resultados individuais aparecem destacados em negrito.	66
Tabela 5.7. Matriz com resultados do teste-t para os valores de Coeficiente de Correlação de Matthews (MCC) com significância de 95%.	67
Tabela 5.8. Valores médios para Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC). Os melhores resultados individuais aparecem destacados em negrito.	68
Tabela 5.9. Matriz com resultados do teste-t para os valores de Coeficiente de Correlação de Matthews (MCC) com significância de 95%.	68
Tabela 5.10. Valores médios para Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC). Os melhores resultados individuais aparecem destacados em negrito.	70
Tabela 5.11. Matriz com resultados do teste-t para os valores de Coeficiente de Correlação de Matthews (MCC) com significância de 95%.	70
Tabela 7.1. Detalhamento de Temas, Conceitos e Palavras-chave utilizadas na construção da base de dados controlada.....	82
Tabela 7.2. Algoritmo utilizado para execução da abordagem sensível ao contexto.	88
Tabela 7.3. Valores médios de Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Área Abaixo da Curva (AUC) nos experimentos <i>Bag-of-visual-features</i> (BoVF), <i>Bag-</i>	

of-visual-differences (BoVD), *Bag-of-topics* (BoT) e *Bag-of-topics-differences* (BoTD), utilizando os descritores SIFT e STIP na base controlada. 94

Tabela 7.4. Matriz com resultados do teste-t par-a-par, com significância de 95%, nos experimentos *Bag-of-visual-features* (BoVF), *Bag-of-visual-differences* (BoVD), *Bag-of-topics* (BoT) e *Bag-of-topics-differences* (BoTD), na base controlada. 95

Tabela 7.5. Valores médios de Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Área Abaixo da Curva (AUC) nos experimentos *Bag-of-visual-features* (BoVF), *Bag-of-visual-differences* (BoVD), *Bag-of-topics* (BoT) e *Bag-of-topics-differences* (BoTD), utilizando os descritores SIFT e STIP na base selvagem. 98

Tabela 7.6. Matriz com resultados do teste-t par-a-par, com significância de 95%, nos experimentos *Bag-of-visual-features* (BoVF), *Bag-of-visual-differences* (BoVD), *Bag-of-topics* (BoT) e *Bag-of-topics-differences* (BoTD), na base selvagem. 99

Lista de Abreviaturas

ANOVA:	Análise de Variâncias (<i>Analysis of Variance</i> , em inglês);
AUC:	Área Abaixo da Curva (<i>Area Under Curve</i> , em inglês);
BoT:	Bag-of-topics;
BoTD:	Bag-of-topic-differences;
BoVD:	Bag-of-visual-differences;
BoVF:	Bag-of-visual-features;
CBIR:	Recuperação de Imagens Baseada no Conteúdo (<i>Content-Based Image Retrieval</i> , inglês);
DoG:	Diferença de gaussianas (<i>Difference-of-Gaussian</i> , em inglês);
FPR:	Taxa de Falsos Positivos (<i>False Positive Rate</i> , em inglês);
GMM:	Modelo de Misturas Gaussianas (<i>Gaussian Mixture Models</i> , em inglês);
HMM:	Modelo Oculto de Markov (<i>Hidden Markov Models</i> , em inglês);
HOF:	Histogramas de Fluxo Óptico (<i>Histograms of Optical Flow</i> , em inglês);
HOG:	Histogramas de Gradientes Orientados (<i>Histograms of Oriented Gradients</i> , em inglês);
IDF:	Frequência de Termos Invertida (<i>Inverse Term Frequencies</i> , em inglês);
LDA:	Alocação Latente de Dirichlet (<i>Latent Dirichlet Allocation</i> , em inglês);
LSA:	Análise de Semântica Latente (<i>Latent Semantic Analysis</i> , em inglês);
LSA:	Análise de Semântica Latente (<i>Latent Semantic Analysis</i> , em inglês);
MCC:	Coeficiente de Correlação de Matthews (Matthews

	Correlation Coefficient, em inglês);
PCA:	Análise de Componentes Principais (<i>Principal Components Analysis</i> , em inglês);
pLSA:	Análise Probabilística de Semântica Latente (Probabilistic Latent Semantic Analysis, em inglês);
SIFT:	Transformada de Características Invariante à Escala (<i>Scale-invariant Feature Transform</i> , em inglês);
STIP:	Pontos de Interesse no Espaço-Tempo (<i>Space-Time Interest Points</i> , em inglês);
SVD:	Decomposição de Valores Singulares (<i>Singular Value Decomposition</i> , em inglês);
SVM:	Máquina de Vetores de Suporte (<i>Support Vector Machines</i> , em inglês);
TF:	Frequência de Termos (<i>Term Frequencies</i> , em inglês);
TPR:	Taxa de Verdadeiros Positivos (<i>True Positive Rate</i> , em inglês);
VSUMM:	Resumo de Vídeo (<i>Video SUMM</i> ary, em inglês);

Sumário

Agradecimentos	iii
Resumo	iv
Abstract	v
Lista de Figuras.....	vi
Lista de Tabelas	vii
Lista de Abreviaturas	ix
Sumário.....	xi
Capítulo 1.....	1
1.1 Motivação.....	3
1.2 Descrição do Problema.....	5
1.3 Objetivos.....	9
1.3.1 Objetivo Geral	9
1.3.2 Objetivos Específicos.....	9
1.4 Hipóteses	10
1.5 Contribuições.....	11
1.6 Organização do Trabalho	12
Capítulo 2.....	14
2.1 Redes Sociais para Compartilhamento de Vídeo	14
2.2 Ações não colaborativas	17
2.3 Considerações.....	19
Capítulo 3.....	21
3.1 Ações não colaborativas em redes sociais para compartilhamento de vídeo...21	
3.1.1 Caracterização de usuários não colaborativos.....	22
3.1.2 Classificação de vídeos com base no volume de acessos	23
3.2 Detecção de <i>spam</i> em diferentes contextos	23

3.3	Identificação de Ações ou Eventos	24
3.4	Classificação de vídeos e imagens em categorias semânticas de âmbito geral	27
3.5	Resumo estático de vídeos	29
3.6	Extratores de características visuais	30
3.6.1	Extratores de características globais	31
3.6.2	Extratores de características locais	33
3.7	Representação de características em nível-médio	36
3.8	Classificação supervisionada	38
3.8.1	SVM	39
3.8.2	Naïve bayes	40
3.9	Combinação de evidências	41
3.10	Adequação do espaço geométrico	42
3.10.1	Análise de Semântica Latente	43
Capítulo 4.....		45
4.1	“Solução modelo” com combinação de evidências	46
4.2	Avaliação de Características Visuais: Atributos Globais × Atributos Locais	49
4.3	Avaliação de Características Atributos Espaço-Temporais × Atributos Espaciais	50
4.4	Avaliação de Características Multimodais: Combinação de Evidências Multimodais × Evidências Monomodais	51
4.5	Avaliação de Características Visuais: Combinação de Evidências × Evidências Isoladas	52
4.6	Avaliação de Classificadores: SVM × Naïve Bayes	54
4.7	Conclusões.....	55
Capítulo 5.....		56
5.1	Aquisição da Base de Dados	56
5.1.1	Base de Dados “Selvagem”	57
5.2	Visão Geral dos Experimentos	59
5.2.1	Experimentos <i>Stuffing</i> × <i>Legítimo</i>	61

5.2.2	Experimentos <i>Spam</i> × Legítimo	62
5.3	Experimentos e Resultados	62
5.4	Avaliando descritores visuais em atuação isolada	63
5.5	Avaliando descritores visuais em atuação conjunta.....	65
5.6	Avaliando classificadores SVM e Naïve Bayes	67
5.7	Avaliando a combinação de evidências multimodais.....	69
5.8	Considerações.....	70
Capítulo 6.....		73
6.1	Representação de Características Visuais Sensível ao Contexto.....	74
6.2	Avaliação de Centralização: Centralização Comum × Centralização baseada na localização contextual.....	76
6.3	Avaliação de Adequação da Geometria: Espaço de Características × Espaço de Semântica Latente	78
Capítulo 7.....		80
7.1	Aquisição da Base de Dados	80
7.1.1	Base de dados controlada	81
7.2	Pseudoalgoritmo para abordagem sensível ao contexto	87
7.3	Visão Geral dos Experimentos	92
7.4	Avaliando a adoção da representação baseada na localização contextual dos vídeos e a transformação linear do espaço de projeção em ambiente controlado ...	93
7.5	Avaliando a adoção da representação baseada na localização contextual dos vídeos e a transformação linear do espaço de projeção em ambiente real.....	97
7.6	Considerações.....	100
Capítulo 8.....		101
8.1	Considerações Finais	101
8.2	Trabalhos Futuros	102
8.3	Trabalhos Publicados	104

Capítulo 1

Introdução

A massificação do acesso em alta velocidade à internet, vivenciado em vários países do mundo, tem gerado uma revolução social comparável às da energia elétrica e telefone; através de um grande número de novos modos de comunicação e prestação de serviços, como blogs, mensagens instantâneas, redes sociais, comércio eletrônico, entre outros. Destes, as redes sociais virtuais [Newman e Park, 2003] são os que têm conseguido maior volume de usuários, em um conjunto impressionante de aplicações.

O surgimento e posterior popularização de mecanismos para produção e distribuição do vídeo digital têm, também, contribuído para mudança no modo como as pessoas se comunicam e interagem. O vídeo digital surgiu comercialmente a partir de 1986, com o formato Sony D-1¹. Inicialmente utilizado, principalmente, por grandes redes de televisão, a tecnologia de aquisição de imagens passou por uma evolução, em um curto período de tempo, com a redução do tamanho dos equipamentos e aumento da qualidade da captura de cenas. Esse avanço permitiu a redução tanto dos custos de produção quanto do preço de equipamentos de captura de vídeo de boa qualidade.

A atuação conjunta desses dois fatores, advento das redes sociais virtuais e o avanço da tecnologia de geração de vídeo digital, permitiu a criação de uma nova forma de interação entre as pessoas, as redes sociais para compartilhamento de vídeo digital. Truong e Venkatesh [2007] atribuem a isso a ampliação significativa do volume de vídeos disponibilizados na web.

Os novos canais de comunicação, criados a partir da internet, trouxeram a necessidade de busca e tratamento da informação em níveis nunca antes imagináveis.

1 *Sony History – The Beginning of the Digital Age*, disponível em:

<http://www.sony.net/SonyInfo/CorporateInfo/History/SonyHistory/2-04.html#block4>, acessado em Agosto de 2012.

Essas necessidades, contudo, ainda estão longe de serem atendidas, sobretudo quando focamos no vídeo digital. Os usuários anseiam por instrumentos que venham a auxiliá-los na recuperação e manipulação do vídeo e que sejam tão eficientes e eficazes quanto os disponíveis para a informação textual.

Incluem-se nessas necessidades mecanismos que permitam coibir abusos e infrações passíveis de serem cometidas pelos usuários desses canais. Entre essas ações estão: instigar a irritação de outros usuários, difundir material inapropriado para os fins da comunidade (por exemplo, propagandas ou pornografia) e manipular ilegitimamente a lista de classificação de popularidade. Esse tipo de comportamento, entendido como “não colaborativo”, polui os canais de comunicação e dificulta às comunidades virtuais alcançarem seus objetivos originais de discussão, aprendizado e entretenimento.

Foi observado por Backstrom et al. [2006] que a permanência dos usuários em redes sociais tem estreita relação com o modo como são tratados os seus tópicos de interesse. Mas, além de implicar a redução da satisfação com o serviço, a ação não colaborativa consome recursos computacionais, que certamente seriam mais bem aproveitados se empregados para fins construtivos.

Seria desejável o provimento de mecanismos automáticos para a detecção e filtragem de ações e conteúdos não colaborativos em redes sociais, de forma a tornar mais eficiente e eficaz a intervenção dos moderadores dessas redes. Entretanto, esses mecanismos automáticos envolvem tarefas complexas de aprendizado e classificação.

Além dessas complexidades, a filtragem de conteúdo enfrenta a enorme quantidade de dados das redes sociais. Um algoritmo automático pode ser um forte aliado para permitir a detecção de conteúdo e ações não colaborativas, mas é crucial que o número de falso-positivos seja mantido baixo, caso contrário, os agentes humanos serão sobrecarregados para a verificação dessas ocorrências.

Tradicionalmente, as soluções para classificação e busca de vídeos envolvem sua associação às palavras-chave [Liu et al., 2007]. Contudo, por mais que a recuperação de documentos textuais esteja, em geral, mais avançada do que a de recuperação de documentos multimídia, a busca textual de documentos visuais está longe de produzir resultado plenamente satisfatório, tanto pelas dificuldades inerentes de se traduzir o conteúdo visual em uma descrição consistente e repetível, quanto pelo enorme custo em

fazer isso manualmente. A essas dificuldades se somam aquelas inerentes ao uso de palavras-chave: sinônimos, generalização, regionalismos, etc. [Yoshitaka e Ichikawa, 1999].

As abordagens que se apresentam mais promissoras são as que têm procurado extrair e atribuir significado ao conteúdo visual presente no vídeo. Essas abordagens se incluem em uma linha de pesquisa conhecida como Recuperação de Imagens Baseada no Conteúdo (do inglês, *Content-Based Image Retrieval* – CBIR) (Kato, 1992). Contudo, a área de CBIR em vídeo possui grandes desafios a serem superados, tais como: (1) o grande volume de dados; e (2) a identificação e representação da informação semântica presente nos vídeos.

Este trabalho se propõe a enfrentar essas dificuldades, apresentando uma alternativa para avançar o estado da arte em classificação e recuperação semântica de vídeos digitais, considerando o seu conteúdo visual. É adotada a extração automática de descritores espaço-temporais, altamente discriminantes, e são empregadas técnicas de aprendizado de máquina, capazes de prover a generalização necessária para a busca de categorias complexas. O cenário de aplicação adotado é a identificação de vídeos não colaborativos em uma rede social virtual.

1.1 Motivação

Redes sociais que se destinam ao compartilhamento de vídeos na web têm vivenciado um grande sucesso: estatísticas demonstram que dezenas de horas de gravação em vídeo são postadas a cada minuto [Burgess e Green, 2009]. A facilidade para gerar e divulgar conteúdo na web aumenta a demanda por ferramentas para estimular sua utilização e, ao mesmo tempo, coibir abusos.

Abordagens tradicionais, utilizando buscas por palavras-chave ou outra descrição textual fornecida com o vídeo, enfrentam diversos problemas, os quais persistem mesmo com a obrigatoriedade do preenchimento de alguns metadados quando da postagem do vídeo. A busca por um vídeo desejado pode ser bastante demorada para o usuário, devido ao grande volume de vídeos e falta de ferramentas eficientes para auxiliar nesse processo [Colombino et al., 2010]. No entanto, as redes sociais para compartilhamento de vídeo provêm outros mecanismos na tentativa de facilitar a busca por um vídeo de interesse,

tais como: lista de mais relevantes, vídeos favoritos, vídeos relacionais, vídeos-resposta, entre outros.

O vídeo-resposta é uma característica de uma das redes sociais mais populares, o YouTube. Nesse mecanismo, um usuário pode postar um vídeo como comentário a outro vídeo. No entanto, essa interessante característica deixa margem para a prática de ações abusivas, como a postagem de respostas não relacionadas com o vídeo original. Esse tipo de ação é chamada de *spamming* e por não contribuir para o bom funcionamento da rede social é entendida como sendo não colaborativa. Esses vídeos podem ou não conter propagandas comerciais; usualmente são postados na expectativa de pegar carona na popularidade de uma determinada discussão ou assunto. Ainda, esse tipo de ação pode ser motivada pelo desejo de instigar respostas agressivas por parte de outros usuários (*trolling*, em inglês).

A definição adotada para *spamming* segue os mesmos princípios adotados por Benevenuto [2009], sendo baseada na semântica dos vídeos em detrimento de sua aparência, considerando o contexto onde o vídeo está inserido. Com essa definição, o contexto tem papel fundamental na sua classificação. É por possuir, ou não, um conteúdo semântico similar ao da discussão em que está inserido, que um vídeo é classificado como sendo legítimo ou *spam*.

O problema de classificação abordado neste trabalho se assemelha ao enfrentado em outras áreas, tais como: a identificação de ações humanas em vídeo [Niebles et al., 2006; Lopes et al., 2009b]; a identificação de pornografia em vídeo [Lin et al., 2003; Lopes et al., 2009a]; entre outros. Contudo, por se tratar da identificação de ações não colaborativas, é necessário analisar fatores externos ao vídeo para avaliar seu contexto e definir sua categoria. Isso se apresenta como um complicador, inviabilizando a adoção ingênua das abordagens propostas nas outras áreas.

Devido ao exposto, o desenvolvimento de uma metodologia capaz de classificar vídeos com base no conteúdo visual em categorias semânticas e, considerando o contexto no qual estão inseridos, apresentou-se como a motivação para que este trabalho tomasse forma. Além disso, o cenário aplicativo escolhido se apresentou como uma oportunidade de avançar em direção à solução de um problema, que causa transtornos às mantenedoras de comunidades virtuais, bem como a seus usuários.

1.2 Descrição do Problema

O problema abordado neste trabalho é o da classificação de vídeos em categorias semânticas utilizando apenas as informações visuais. O que o diferencia em relação a outras referências da literatura, é o fato das categorias serem dependentes da interpretação semântica de outros vídeos ligados ao contexto do vídeo sendo classificado.

A classificação em categorias semânticas para imagens e vídeos é um dos maiores desafios da visão computacional, como observado por Snoek e Worring [2009]. Isso se deve ao fato de as categorias em que os objetos serão classificados estarem relacionadas a conceitos complexos, com enorme variabilidade visual, trazendo à tona o problema do *semantic gap* (fosso semântico, em português).

A existência do fosso semântico, que é a divergência entre a informação de baixo-nível extraída dos documentos e a interpretação de alto-nível desses mesmos documentos pelos usuários [Smeulders et. al., 2000], se apresenta como a grande barreira à busca e classificação baseadas no conteúdo visual. De fato, as técnicas existentes para essas tarefas ainda estão longe de atingir o nível de acurácia de um operador humano.

Situações em que é necessária a diferenciação entre categorias com características visuais marcantes e distintas (como um jogo de futebol *versus* um desenho animado, por exemplo) já demonstraram bons resultados. O problema abordado neste trabalho é a elaboração de uma abordagem capaz de distinguir, com um grau aceitável de qualidade, entre elementos pertencentes a categorias extremamente complexas. Aqui, as características visuais apresentam enorme variabilidade intraclasse e a definição da classe a qual o vídeo pertence é diretamente dependente da semântica da discussão na qual está posicionado.

Devido à complexidade do problema abordado, este trabalho se concentra em problemas envolvendo apenas duas classes, o que simplifica consideravelmente o aparato conceitual e estatístico da classificação. Essa restrição, porém, não é tão severa quanto parece a princípio, e muitas aplicações úteis se enquadram nela, tal como o clássico problema da detecção de *spam* em *e-mails*.

No caso deste trabalho, a aplicação escolhida para a avaliação da metodologia proposta é a identificação da presença de vídeos-resposta que possuem conteúdo com semântica divergente ao tema do vídeo original a que foi associado. Mais precisamente, foi desenvolvida uma ferramenta para identificar a ocorrência de vídeos-resposta do tipo *stuffing* e *spam* (definidos a seguir) dentre um conjunto de vídeos-resposta. Esses vídeos são entendidos no contexto deste trabalho como sendo ações não colaborativas para o bom comportamento da rede social. Para a realização desta tarefa, foram analisados e comparados apenas o conteúdo visual presente nos vídeos e o contexto no qual estão inseridos.

Os vídeos-resposta não colaborativos são ações de usuários que, com intenção ou não, prejudicam o bom funcionamento das redes sociais. Esse tipo de ação pode ocorrer de diferentes formas. Neste trabalho, um vídeo não colaborativo é compreendido como sendo um vídeo-resposta com tema divergente do conteúdo do vídeo original. Para os fins deste trabalho, são considerados dois casos:

- Vídeos com conteúdo não relacionado ao do vídeo original, que possuem curta duração, grande ocorrência de repetição de conteúdo visual em seus quadros e são similares a outros vídeos na mesma *thread*² de vídeos-resposta, possivelmente variando apenas a cor ou posicionamento dos objetos. Esse tipo de vídeo pode ser chamado também de *ballot stuffing* (“fraude de enchimento de urna”, o ato de colocar mais de um voto na urna eleitoral), ou apenas *stuffing*, pois tem o objetivo de promover, ou ampliar a visualização de um dado vídeo original;

2 *Thread*, é uma palavra da língua inglesa que significa “linha”. Nesta situação em específico, é interpretada como sendo uma “linha de discussão”. Em um fórum, uma discussão pode gerar respostas, e essas podem receber respostas (resposta da resposta). Ou seja, *thread* é o termo utilizado para definir uma cadeia de respostas em uma discussão.

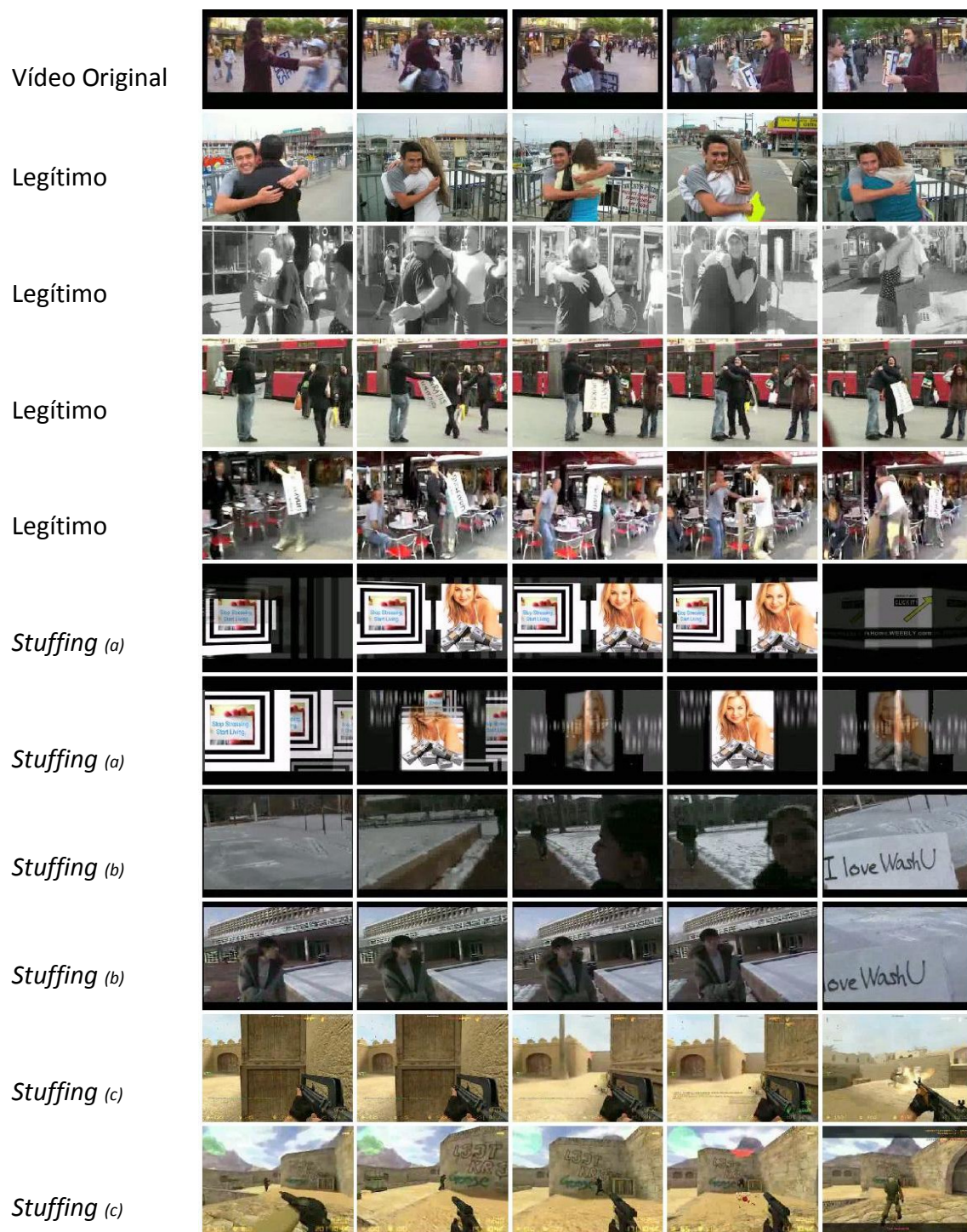


Figura 1.1. Quadros extraídos de vídeos de uma *thread* de vídeos-resposta exemplificando a ocorrência de elementos da categoria *Stuffing*. Os vídeos que compartilham conteúdo estão identificados com uma letra entre parênteses.

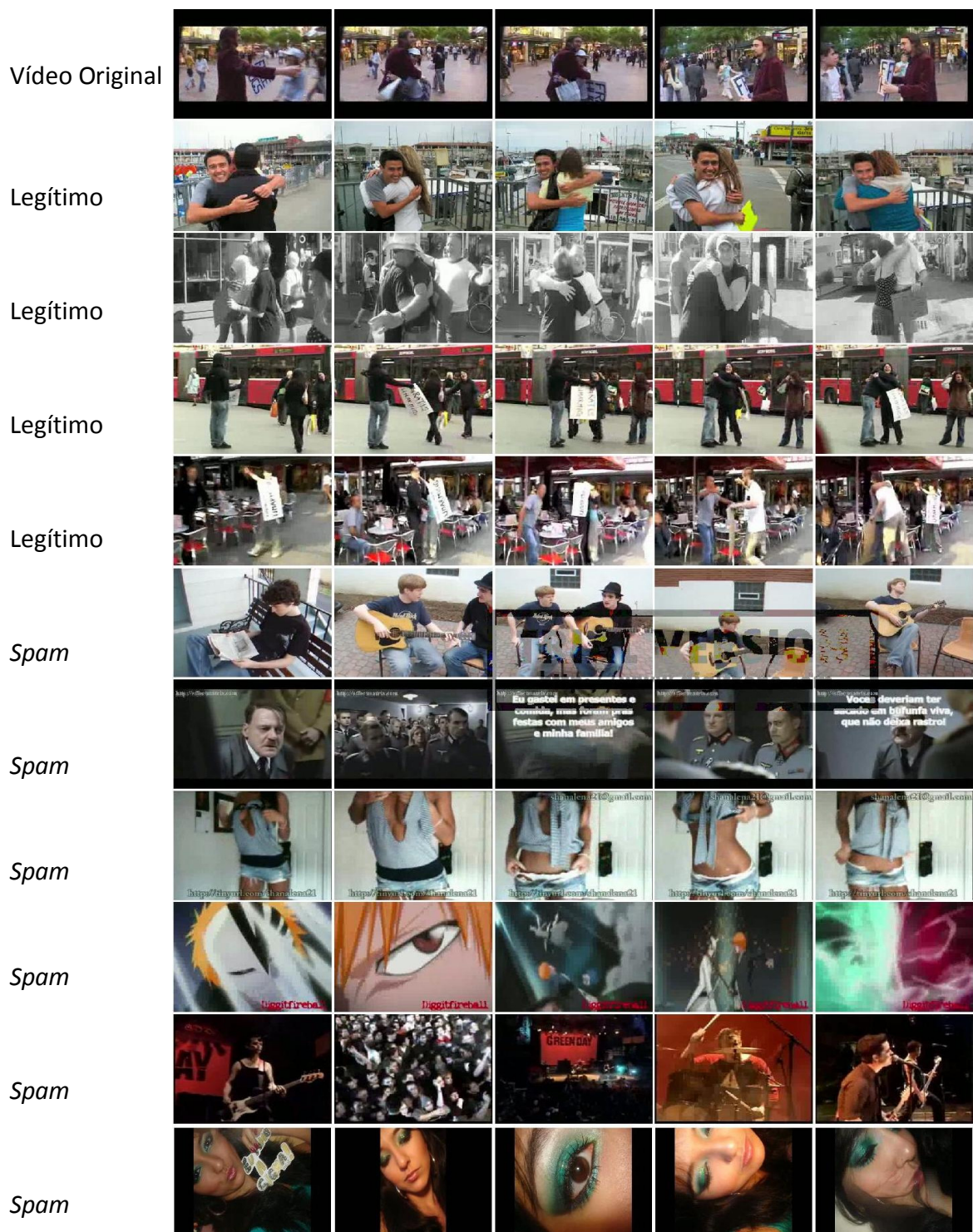


Figura 1.2. Quadros extraídos de vídeos de uma *thread* de vídeos-resposta exemplificando a ocorrência de elementos da categoria *Spam*.

- Vídeos que possuem conteúdo tão rico quanto os legítimos, mas sua informação não está relacionada à temática do vídeo original. Para caracterizar um vídeo como pertencente a esta categoria é necessário

contrastar a sua semântica com a semântica da *thread* em que ele se encontra. Os vídeos pertencentes a esta categoria podem ser chamados de *spam*, aproveitando a mesma nomenclatura utilizada para classificar mensagens indesejadas enviadas a usuários de correio eletrônico.

As Figuras 1.1 e 1.2 apresentam exemplos de quadros extraídos de vídeos que se enquadram em cada um dos casos acima. Para cada figura são apresentados o vídeo original, um conjunto de respostas legítimas e exemplos de vídeos não colaborativos.

1.3 Objetivos

1.3.1 Objetivo Geral

Este trabalho tem como objetivo geral avançar o estado da arte da detecção de conteúdo não colaborativo em redes sociais através da melhor representação de características visuais, incorporando as informações de contexto dos vídeos.

1.3.2 Objetivos Específicos

Os objetivos específicos definidos neste trabalho são:

- Analisar o desempenho de características visuais de baixo-nível, considerando diversas possibilidades de escolha, como: descritores locais *versus* globais; puramente espaciais *versus* espaço-temporais, e; adição de informação extra sobre o vídeo (comprimento e variabilidade visual). Na Seção 3.6, são apresentadas descrições dos extratores de características de baixo-nível utilizados;
- Contrastar o desempenho de dois classificadores tradicionais para a classificação de imagens, o SVM (discriminante) e o Naïve Bayes (generativo). Na Seção 3.8, são apresentadas as descrições das diferentes abordagens de classificação avaliadas;
- Analisar o desempenho da utilização das evidências visuais isoladas, em contraste ao uso dessas evidências combinadas em *late fusion* (após a

classificação). Na Seção 3.9, é apresentada descrição da estratégia de combinação de evidências adotada;

- Analisar a efetividade do uso de uma representação que incorpora a informação de contexto da vizinhança dos vídeos ao descritor de nível-médio tradicional (*bags-of-visual-features*). Na Seção 3.7, é apresentada a abordagem de representação em nível-médio mais adotada na literatura recente, *bag-of-visual-features*;
- Analisar o desempenho da representação sensível ao contexto em diferentes espaços de dados (original e retificado por uma transformação linear). Na Seção 3.10, são apresentadas características de uma estratégia linear de adequação do espaço geométrico;

1.4 Hipóteses

As hipóteses gerais que nortearam este trabalho são:

- **Hipótese 1 - *Descrição visual de baixo-nível*:** a escolha do descritor visual de baixo-nível terá impacto positivo na acurácia da classificação, sendo contrastados descritores locais *versus* globais; puramente espaciais *versus* espaço-temporais, e; adição de informação extra sobre o vídeo (comprimento e variabilidade visual). A avaliação desta hipótese está dividida nas Seções 5.4 e 5.7;
- **Hipótese 2 - *Abordagem de classificação*:** o classificador SVM, por ser discriminativo e por não considerar as dimensões da representação de nível-médio como independentes, terá melhor desempenho em relação ao Naïve Bayes, que é generativo e assume independência. Esta hipótese é avaliada na Seção 5.6;
- **Hipótese 3 - *Combinação de evidências visuais*:** a utilização conjunta de diferentes evidências visuais contribui positivamente para o aumento da qualidade dos resultados da classificação, se comparada ao uso isolado dessas evidências. Esta hipótese é avaliada na Seção 5.5;

- **Hipótese 4 - *Representação de nível-médio*:** uma representação que incorpora o contexto no qual o vídeo está sendo analisado proporciona maior qualidade nos resultados da classificação em comparação ao uso da representação de *bag-of-visual-features* tradicional. A avaliação desta hipótese está dividida nas Seções 7.3 e 7.4, considerando o desempenho em ambientes controlado e real, respectivamente;
- **Hipótese 5 - *Transformação do espaço*:** a transformação do espaço de projeção do vetor de características de nível-médio torna as distâncias entre vídeos relacionados mais uniformes em todo o espaço, melhorando a representação sensível ao contexto em contraste com a projeção no espaço original. A avaliação desta hipótese está dividida nas Seções 7.3 e 7.4, considerando o desempenho em ambientes controlado e real, respectivamente.

Consideramos como melhoria significativa quando é obtido um incremento estatisticamente significante na métrica de interesse (por exemplo, o Coeficiente de Correlação de Matthews).

1.5 Contribuições

A detecção automática de comportamentos não colaborativos em redes sociais para o compartilhamento de vídeo teve como pioneiros Benevenuto et al. [2009], que se propuseram a detectar usuários com comportamento não colaborativo (*spammers*) utilizando informação extraída do perfil do usuário, relacionamentos na rede social e a frequência de postagens. Embora o contexto de aplicação deste trabalho seja o mesmo que o realizado por Benevenuto et al. [2009], é seguido um caminho diferente em dois importantes aspectos.

Primeiramente, o interesse se concentra em caracterizar **ações** ao invés de **usuários**. Ou seja, nesta abordagem realiza-se a classificação de cada um dos vídeos-resposta, ao invés de tentar caracterizar o comportamento dos usuários que os postaram. É considerado que existem diferentes graus de comportamentos não colaborativos e que mesmo um *spammer* também pode vir a postar vídeos-resposta legítimos.

A segunda diferença importante são as características escolhidas para representar os vídeos. Elegeram-se o conteúdo visual de forma quase exclusiva. Acredita-se que a presente abordagem é a primeira que se propõe a detectar vídeos não colaborativos utilizando análise baseada em conteúdo visual [Luz et al., 2011]. É preciso frisar que acredita-se que todas as informações disponíveis (metadados, tags, interações sociais, conteúdo visual e trechos de áudio) são relevantes para se construir uma ferramenta mais efetiva no combate às ações não colaborativas. Contudo, neste trabalho, o interesse é em avaliar o quanto a informação visual é útil para realizar essa tarefa.

As principais contribuições apresentadas neste trabalho são:

- a construção de uma abordagem para representação de características visuais capaz de incorporar a informação de contexto dos vídeos, permitindo a classificação dos vídeos do tipo *spam* de diferentes linhas de discussão de forma generalizada. Essa contribuição tem potencial impacto além do presente trabalho, podendo ser usada em diferentes aplicações de classificação de vídeos dependentes do contexto, como sugestão de categorias, correção automática de erros de classificação e anotação, *spam* em outros tipos de redes sociais multimídia (por exemplo, Flickr ou devianART), entre outras;
- baseado nessa representação, um esquema para detecção de *spam* em *threads* de vídeo-resposta, que oferece recursos poderosos de generalização compensando a escassez de conjuntos de treinamento, não requerendo que o conjunto de treinamento possua elementos de cada um dos conceitos a serem analisados;
- foco na análise da informação visual, fornecendo evidências robustas, já que não dependem da corretude ou disponibilidade dos metadados.

1.6 Organização do Trabalho

Este trabalho está dividido em três partes principais. Na primeira parte, compreendida pelos Capítulos 2 e 3, são apresentados o estado da arte na área de pesquisa e a

fundamentação teórica das abordagens utilizadas. Na segunda e terceira parte, apresentam-se as contribuições da presente tese.

O trabalho se estruturou em dois momentos distintos. No primeiro, houve principalmente preocupação com a extração de características de baixo-nível (cor, textura, momentos invariantes, características locais baseadas em pontos de interesse, etc.), avaliando, dentre as técnicas disponíveis na literatura, quais as mais promissoras. Também se apresentou o interesse em avaliar diferentes formas de utilizar e combinar essas características. Os resultados dessa avaliação são explorados na segunda parte do trabalho, compreendendo os Capítulos 4 e 5.

No segundo momento, explorou-se a composição das características chamadas de “nível-médio” (*bags-of-features* e seus variantes). Aqui, já estava clara a necessidade de incorporar a informação de contexto à representação, para que o classificador não precisasse levar em conta esse contexto, usando mecanismos explícitos e *ad hoc*. A terceira parte da tese, que compreende os Capítulos 6 e 7, apresenta essa metodologia de representação sensível ao contexto.

Por fim, no Capítulo 8, são apresentadas algumas considerações finais.

Capítulo 2

Caracterizando atitudes não colaborativas

Neste capítulo, são apresentadas algumas características de redes sociais para compartilhamento de vídeo, na Seção 2.1. Além disto, são caracterizadas as ações não colaborativas de interesse para o trabalho, na Seção 2.2.

2.1 Redes Sociais para Compartilhamento de Vídeo

Newman e Park [2003] caracterizam as redes sociais pela presença de um grupo de pessoas com interesses comuns que se relacionam entre si. Cada vez mais as pessoas têm se conectado através deste tipo de rede com as mais diversas finalidades, desde o compartilhamento de opiniões (por exemplo, *LiveJournal*³ e *Twitter*⁴), imagens (por exemplo, *Flickr*⁵) ou vídeos (por exemplo, *YouTube*⁶), passando por identificação de oportunidades de negócio (por exemplo, *LinkedIn*⁷), busca de relacionamentos (por exemplo, *MySpace*⁸ e *Orkut*⁹), dentre outras.

3 www.livejournal.com

4 twitter.com

5 www.flickr.com

6 www.youtube.com

7 www.linkedin.com

8 www.myspace.com

9 www.orkut.com

Com o avanço da tecnologia digital para captura de imagens e vídeos e a massificação do uso da internet, o surgimento das redes sociais virtuais para compartilhamento de vídeo foi uma consequência, senão prevista, inevitável. Em uma rede social virtual, os diversos participantes possuem objetivos e interesses comuns e trocam informações acerca desses conteúdos. Nessas redes, as informações trocadas são, prioritariamente, vídeos digitais com conteúdo representativo de um dado tema de interesse.

Segundo Benevenuto [2009] uma rede social para compartilhamento de vídeo possui quatro características importantes: **a)** usuários podem contribuir com algum tipo de vídeo, frequentemente criado pelo próprio usuário, e indicar um título, descrição e um conjunto de palavras-chave para este; **b)** usuários podem visualizar o conteúdo de vídeos postados por outros usuários; **c)** usuários podem avaliar o conteúdo dos vídeos utilizando um sistema particular a cada Rede, podendo incluir: notas, comentários textuais, vídeos-resposta ou combinações destes, e; **d)** usuários podem manter uma lista de favoritos, que pode incluir: vídeos, usuários, grupos temáticos, entre outros, de acordo com a particularidade da Rede.

Atualmente existem vários sites contendo redes para compartilhamento de vídeo: a Tabela 2.1 apresenta uma relação dos mais populares, segundo o serviço de monitoramento de tráfego na web Alexa¹⁰. Na tabela, os três primeiros colocados são especializados no compartilhamento de vídeos na web e os outros são sites que oferecem mais serviços além do compartilhamento de vídeo e por isso não podem ser comparados diretamente.

O YouTube aparece em primeiro lugar no volume de visitas, sendo que 33,15% de todos os usuários da web realizam visitas a essa comunidade por dia, com uma larga diferença à frente dos outros concorrentes (o Dailymotion, segundo colocado, obtém visitas de 0,89% do total de usuários da rede). Um dos fatores do sucesso do YouTube é o grande conjunto de ações disponibilizadas aos usuários, por exemplo, a possibilidade de publicação de vídeos-resposta a um dado vídeo postado por outro usuário. Um estudo mais aprofundado sobre as características do YouTube foi elaborado por Cha et al. [2007].

10 Disponível em: www.alexa.com, acessado em janeiro de 2012.

Tabela 2.1. Relação de Redes Sociais para compartilhamento de vídeo mais populares¹.

Comunidades	Usuários (%)	Páginas visitas por usuário	Tempo no site (min:seg)
YouTube	33,15	12,73	17:25
Dailymotion	0,89	4,3	5:22
Metacafe	0,29	3,6	3:37
Yahoo! ²	22,79	6,65	9:09
MySpace ³	0,72	2,93	2:57

¹ Fonte: www.alexa.com, em janeiro de 2012.

² Os valores pertencentes ao *Yahoo!* são referentes ao site www.yahoo.com, segundo a fonte, apenas 0,1% do total de acessos a esse site são direcionados ao subdomínio video.yahoo.com que representa o serviço de compartilhamento de vídeo.

³ Os valores pertencentes ao *MySpace* são referentes ao site www.myspace.com, segundo a fonte, apenas 0,5% do total de acessos a esse site são direcionados ao subdomínio vids.myspace.com que representa o serviço de compartilhamento de vídeo.

Os valores de páginas visitadas por usuário e tempo no site auxiliam na identificação da relevância do YouTube no cenário de redes de compartilhamento de vídeo na web. Uma análise mais detalhada das estatísticas de acesso em cada uma dessas comunidades pode ser realizada observando as informações apresentadas na Tabela 2.1:

1. Os usuários do YouTube permanecem nessa comunidade por um tempo muito superior ao permanecido por usuários em outras comunidades, sendo 17:25 min. contra 5:22 min. no Dailymotion, segunda colocada. Essa situação analisada de modo isolado poderia levar a uma conclusão equivocada, visto que o maior tempo de permanência pode ser justificado por usuários assistirem a vídeos com maior tempo de duração no YouTube em relação aos demais, ou mesmo, devido a um maior tempo de resposta no atendimento às solicitações dos usuários;
2. A informação de páginas vistas por usuário representa uma média de quantitativo de páginas que cada usuário do serviço visualiza a cada vez que acessa a comunidade. Neste quesito, o YouTube também aparece na frente dos demais com 12,73 páginas/usuário, enquanto que o segundo colocado apresenta uma relação de 4,3 páginas/usuário (Dailymotion). A simples observação desse fato não permite identificar o YouTube como sendo o serviço mais aceito pelos usuários dentre as redes de compartilhamento de

vídeo. O maior número de páginas visualizadas por cada usuário pode suscitar a dúvida quanto à satisfação do usuário em relação ao conteúdo visualizado na página, o que o faria buscar por mais páginas até localizar a informação desejada.

Em uma análise integrada das informações de volume global de visitas, páginas visitadas por usuário e tempo no site, indicando o YouTube como líder em todas, é possível concluir, com maior segurança, que essa comunidade dispõe do serviço de compartilhamento de vídeo mais utilizado atualmente. A importância do YouTube manifesta pela grande aceitação dos usuários, foi fator motivador ao estudo do mecanismo inovador de *threads* de vídeo-resposta proposto por essa comunidade. Acredita-se que interações desse tipo, diretamente em multimídia rica (áudio, vídeo, imagens, etc.), serão cada vez mais frequentes na web, ao lado da interatividade mais convencional usando votos e comentários textuais.

2.2 Ações não colaborativas

Uma das características presentes nas redes sociais, e que garante a sua boa aceitação por parte dos usuários, é a permissividade presente neste tipo de rede. Ou seja, ao usuário é dada a ampla liberdade de executar suas ações, cabendo a cada membro da rede ter a consciência de agir de forma a não prejudicar o grupo. Contudo, esse autopolicimento nem sempre é praticado por todos os usuários. Quando esses realizam ações que prejudicam outros membros ou, que estejam em desacordo com a conduta esperada pela comunidade, classificamos essas ações como **não colaborativas**.

Todos os tipos de ação não colaborativa acarretam complicações para o mantenedor da rede social e seus membros, tais como:

- uso de espaço em disco desnecessário;
- sobrecarga da banda de rede disponível para a exibição de conteúdo não interessante aos usuários;
- elaboração de consultas com conteúdo incorreto;
- impaciência, insatisfação e descrédito com os serviços oferecidos.

Ações não colaborativas são presenciadas na maioria dos serviços baseados na *web*, tais como: *e-mail*, *blogs*, sites de busca, compartilhamento de imagens e vídeos, enciclopédia digital, comércio eletrônico, entre outros. Para cada nova atitude não colaborativa que surge, uma série de trabalhos são desenvolvidos no intuito de combatê-las [Fetterly et al., 2004; Gomes et al., 2007; Castillo et al., 2007; Gyöngyi et al., 2004; Lin et al., 2008; Thomason, 2007; Wu et al., 2005; Benevenuto et al., 2009; Heymann et al., 2007; Jindal e Liu, 2008; Yu et al., 2006].

Tratando mais especificamente de atitudes não colaborativas em redes de compartilhamento de vídeo, Benevenuto et al. [2009] apresentam diferentes elementos que podem ser alvo deste tipo de ação, tais como:

- **Metadados:** conjunto de informações cadastradas pelo usuário quando posta um vídeo na comunidade virtual para auxiliar na descrição do conteúdo do vídeo e podem ser utilizadas como referência para tarefas como a busca por vídeos com conteúdos similares. A análise de metadados tende a facilitar o processo de busca por vídeos com uma mesma temática, contudo há a possibilidade de usuários mal intencionados cadastrarem uma descrição divergente ao conteúdo do vídeo postado, induzindo erros de classificação na busca de vídeo com similaridade temática;
- **Interface:** as principais redes sociais para compartilhamento de vídeo geralmente possuem uma interface adaptável, exibindo uma lista de melhores (*Top List*, em inglês) vídeos dentre os temas de maior interesse do usuário, e ações de manipulação que impactam na inclusão de vídeos de pouca expressão ou não relacionados nessa lista, são entendidas como não colaborativas. O YouTube, por exemplo, utiliza a quantidade de vídeos-resposta associados a um dado vídeo como um dos critérios para construção de *Top Lists*, possibilitando a usuários mal-intencionados se aproveitarem postando um grande volume de vídeos-resposta, mesmo que todos sejam iguais e não relacionados ao vídeo-original, fazendo com que esse vídeo-original seja incluído na lista;
- **Interações:** a possibilidade de interação direta entre os membros da comunidade é uma das principais características das redes sociais. Além da troca de mensagens textuais através de comentários deixados nos *posts* de outros usuários, existem

pelo menos duas outras possibilidades, cada uma em uma rede em particular, vídeos-resposta (no YouTube) e edição de metadados (no Metacafe). Em cada uma dessas situações é possível a ocorrência de ações não colaborativas, tais como: a postagem de comentários com a intenção de denegrir ou irritar outro membro; o envio de um vídeo-resposta com conteúdo divergente ao do vídeo-original, motivado, dentre outras possibilidades, a divulgar conteúdo pornográfico ou comercial se aproveitando da popularidade de um dado vídeo, e; propositalmente modificar os metadados de um vídeo adicionando informações não relacionadas, fazendo com que deixe de ser associado ao seu tema real;

- **Conteúdo:** as ações nesse elemento podem acontecer de diferentes formas, geralmente para se determinar se um dado vídeo está, ou não, colaborando com a comunidade deve ser analisado o contexto no qual está inserido. Dada a liberdade que cada usuário tem para criar e enviar vídeos para a rede, este se apresenta como o maior espaço para realização de ações não colaborativas. As possibilidades são tantas que seria difícil lista-las de modo exaustivo. São exemplos: **a)** envio de duplicatas de vídeos, ou seja, vídeos com conteúdo que já foram postados anteriormente; **b)** postagem de conteúdo comercial em meio a vídeos comuns; **c)** submissão de vídeos protegidos por direitos autorais de modo não autorizado, e; **d)** veiculação de conteúdo pornográfico em ambiente não destinado a esse fim.

A divisão de elementos elaborada por Benevenuto et al. [2009] não tem por objetivo classificar as diferentes ações não colaborativas em conjuntos isolados. É totalmente possível que uma mesma ação impacte em diferentes elementos. Um exemplo são as ações não colaborativas alvo de investigação neste trabalho, *stuffing* e *spam*, que podem trazer prejuízos para os elementos de Interface, Interação e Conteúdo, ao mesmo tempo. Essa abrangência de repercussão é um dos motivos que fez com que essas ações fossem selecionadas para investigação neste trabalho.

2.3 Considerações

Nesse capítulo foram apresentados conceitos de redes sociais para compartilhamento de vídeo. Pode-se observar que o YouTube se apresenta como a comunidade virtual para

compartilhamento de vídeo mais relevante atualmente, recebendo visitas de mais de 30% do total de usuários da *web*. Tal fato atrai o interesse de se realizar experimento com a abordagem proposta em problemas vivenciados nessa comunidade.

Além disso, foram apresentados diferentes elementos passíveis de sofrerem com as ações não colaborativas executadas por membros da comunidade virtual. E o fato de causarem impacto em mais de um desses elementos, reforçou a seleção das ações do tipo *stuffing* e *spam* como foco de aplicação da abordagem desenvolvida nesse trabalho.

No próximo capítulo são apresentadas referências de literatura para as ferramentas utilizadas na construção da metodologia elaborada, bem como descritos trabalhos recentes que se propõem a abordar problemas similares ao enfrentado neste trabalho, no intuito de contextualizar a relevância e atualidade do tema pesquisado.

Capítulo 3

Revisão de Literatura

Neste capítulo são apresentados trabalhos recentes em áreas de aplicação correlatas ao problema abordado, bem como descritas as técnicas que dão suporte ao desenvolvimento da metodologia elaborada. A detecção de ações não colaborativas em redes sociais para compartilhamento de vídeo é um tema recente, com uma literatura restrita, que é apresentada na Seção 3.1. Também são apresentadas abordagens que atuam em áreas correlatas divididas em: a) detecção de *spam* em diferentes contextos, Seção 3.2; b) identificação de ações ou eventos, Seção 3.3, e; c) classificação de imagens e vídeos em categorias semânticas de âmbito geral, Seção 3.4. As Seções 3.5 a 3.10 apresentam uma descrição sucinta do ferramental técnico utilizado neste trabalho. Nessas seções “utilitárias” o nível de aprofundamento é apenas suficiente para a compreensão da metodologia elaborada, não tendo como objetivo ser uma referência exaustiva do estado da arte dessas ferramentas. A Seção 3.7 recebe mais ênfase e detalhamento, visto apresentar a abordagem de *bag-of-visual-features* que é base para a metodologia elaborada neste trabalho.

3.1 Ações não colaborativas em redes sociais para compartilhamento de vídeo

Devido à expansão do uso de redes sociais virtuais para vídeo ser recente, o desenvolvimento de técnicas para filtragem de conteúdo irrelevante despertou até agora poucas iniciativas de pesquisa. O estado da arte se resume aos trabalhos pioneiros de Crane e Sornette [2008], Benevenuto et al. [2009] e Langbehn et al. [2010]. Contudo, nenhuma das propostas considera o conteúdo visual dos vídeos.

Devido à escassez da literatura e à ausência de outros trabalhos que utilizem o conteúdo visual, não é possível traçar um comparativo direto entre a abordagem desenvolvida neste trabalho com outras já existentes. Contudo, são descritas a seguir as principais características e resultados obtidos nos trabalhos que se propõem a reduzir a incidência de conteúdo não colaborativo em redes sociais para compartilhamento de vídeo.

3.1.1 Caracterização de usuários não colaborativos

O foco de atuação dos trabalhos desenvolvidos nesta linha é a identificação de usuários que realizam ações não colaborativas. Os vídeos são classificados como não colaborativos como consequência do comportamento dos usuários que os postaram. Essa abordagem foi inicialmente proposta por Benevenuto et. al. [2009], utilizando características extraídas dos perfis de usuários, relacionamentos e frequência de conteúdo postado nas redes sociais, aliadas ao uso do classificador SVM.

Uma extensão ao trabalho original foi apresentada por Langbehn et al. [2010], aliando a utilização de um classificador multivisão ao uso de regras de associação. Com essa extensão foi possível reduzir o tamanho dos conjuntos de treinamento, sem comprometer a qualidade dos resultados.

Em ambas as iniciativas, os usuários são classificados em “legítimos”, “*promoters*” e “*spammers*”. Essas denominações inspiram as adotadas no presente trabalho: um usuário do tipo “*promoter*” é aquele que envia vários vídeos similares e com conteúdo irrelevante em uma mesma *thread* de vídeo-resposta; um usuário do tipo “*spammer*” é aquele que divulga um vídeo que não tem relação ao tema em discussão nas *threads*, e; os usuários do tipo “legítimo” são os que postam apenas vídeos com informações pertinentes às temáticas das *threads*.

Os resultados obtidos por Benevenuto et. al. [2009, 2012] e Langbehn et al. [2010] alcançam uma taxa elevada de acertos na identificação de usuários do tipo *promoter* (em torno de 92% para taxa de verdadeiros positivos e 0,5% para taxa de falsos positivos). Porém, os resultados da identificação de *spammers* não são tão satisfatórios (em torno de 58% para taxa de verdadeiros positivos e 4% para taxa de falsos positivos), indicando oportunidade para outras abordagens complementares, como a presente tese.

3.1.2 Classificação de vídeos com base no volume de acessos

Nesta linha de atuação o objetivo é a classificação de vídeos de acordo com a sua contribuição na rede social, analisando a probabilidade de recorrência de eventos, considerando como critério a probabilidade de um dado vídeo ser acessado múltiplas vezes. Para isso é analisada a “distribuição de *Poison*” [FA, 1967].

A proposta se baseia na verificação de que 90% dos vídeos são pouco visualizados e apenas 10% recebem maior foco [Crane et al., 2008]. Os vídeos são classificados em três classes: *junk*, vídeos que recebem um bom volume de acessos, mas não são repassados a outros usuários, geralmente por não possuir conteúdo relevante; *quality*, vídeos que também recebem um bom volume de acessos, e são repassados para outros usuários, e; *viral*, vídeos que recebem um volume muito grande de acessos, em um curto espaço de tempo, independentemente de serem repassados para outros usuários.

Segundo observado por Crane e Sornette [2008], essa proposta obteve resultados satisfatórios na identificação de elementos do tipo *viral*, contudo falha na tentativa de separação entre vídeos *quality* e *junk*. Uma das razões observadas é o pouco volume de dados de treinamento.

3.2 Detecção de *spam* em diferentes contextos

Existem muitas abordagens que enfrentam o problema de detecção de comportamentos não colaborativos do tipo *spam*, em uma variedade de cenários. Nesta seção, é apresentado um resumo de alguns desses esforços, primando por aqueles que possuem proximidade com o problema abordado neste trabalho.

Um cenário bem conhecido é o do *e-mail*. O objetivo é a identificação de características no tráfego ou conteúdo das mensagens que possibilitem a distinção entre *e-mails* legítimos e *spam*. Os trabalhos de Drucker et al. [1999] e Androutsopoulos et al. [2000] são alguns dos precursores dessa área. Mais recentemente, Koprinska et al. [2007] e Lai et al. [2009], propuseram técnicas automáticas, baseadas em análise do conteúdo textual, para identificação de *e-mails* maliciosos. Adicionalmente, Wu e Tsai [2009]

apresentaram uma abordagem para caracterização do comportamento dos *spammers* como alternativa para filtragem do conteúdo não colaborativo.

Recentemente a tarefa de detecção de *spam* em *e-mails* recebeu novos desafios. Os *spammers* adaptaram suas táticas introduzindo o chamado *image spam* (em inglês). Essa tática consiste em utilizar uma imagem contendo o texto da mensagem a ser difundida e enviá-la como anexo do *e-mail*. Com isso, as técnicas tradicionais baseadas em análise textual são dificultadas. Vários trabalhos se propõem a enfrentar esse problema, tal como o apresentado por Mehta et al. [2008] que utiliza as informações de cor, com base nos canais do espaço *Lab*, textura, anisotropia e contraste, e localização espacial dos *pixels*, obtida com as coordenadas *x* e *y*, associadas ao uso do Modelo de Misturas Gaussianas para construir uma estratégia de detecção de *image spam* baseada na detecção de duplicatas, considerando as imagens construídas de acordo com um *template*. A revisão elaborada por Biggio et al. [2011] apresenta uma avaliação experimental de várias abordagens de detecção de *image spam*.

A ocorrência de perfis falsos e conteúdo indesejável são comuns em redes sociais. Wang [2010] aborda esse problema com foco na detecção de *spam bots* no Twitter, propondo uma técnica baseada em aprendizado de máquina, características da rede social (tais como o número de amigos e o número de seguidores do usuário) e, também, as 20 postagens mais recentes do usuário. Já Lee et al. [2010] apresentam uma abordagem, denominada *Social Honeypot Project* (em inglês), que utiliza o Twitter e o MySpace como ambientes experimentais para detecção de *spammers* em redes sociais. São coletados perfis de alguns *spammers* e construído um classificador com base na análise estatística das propriedades desses perfis. Irani et al. [2010] também realizaram um comparativo de algumas abordagens que se propõem realizar a detecção de *spam bots*.

3.3 Identificação de Ações ou Eventos

A identificação de ações ou eventos em vídeos é recorrentemente abordada pelas áreas de visão computacional e aprendizado de máquina. As abordagens variam de acordo com o foco de aplicação, variando desde a identificação de texturas dinâmicas, como água ou fogo, à identificação de ações isoladas, como pessoas sorrindo ou sentando/levantando.

Como toda tarefa de classificação visual, a identificação de ações sofre os efeitos do fosso semântico, além de enfrentar as grandes variações de iluminação, presença de oclusões e deformações nos objetos de interesse. Muitos métodos de classificação de ações tem se interessado pelo uso de informações de contexto locais (influência do tipo de cena, ou do tipo de gênero de vídeo na probabilidade das diferentes ações, por exemplo). Entretanto, esses métodos ainda não exploram o tipo de dificuldade que enfrentamos no presente trabalho que é a dependência de um contexto composto por outros vídeos para a classificação do vídeo de interesse (que se encontra, sobretudo no caso do *spam*).

A metodologia para identificação de ações humanas em cenas de filmes de cinema, elaborada por Laptev et al. [2008], visa resolver o problema de anotação das cenas combinando informações textuais, obtidas de legendas e roteiros, e informação de tempo de ocorrência da cena para construir uma ferramenta de anotação automática. A identificação automática das ações é realizada com base em informação visual extraída com técnicas espaço-temporais e utilizando um classificador SVM. Os resultados experimentais demonstraram resultados superiores a demais técnicas da literatura.

No contexto de detecção de ações e eventos, Lopes et al. [2009b, 2009c] propõem o uso associado de descritor espaço-temporal de baixo-nível e representação de nível-médio. Enquanto Souza et al. [2010] aplicam esse *framework* no contexto da detecção de vídeos com cenas de violência. Ambos os experimentos demonstraram resultados superiores ao uso de descritores puramente espaciais.

No trabalho elaborado por Zhang et al. [2009] é descrito um novo *framework* para o aprendizado de modelos de cenas baseando-se em conceitos semânticos. É realizada uma detecção prévia dos objetos de acordo com duas categorias, veículos ou pedestres. Em seguida, é realizado o aprendizado dos padrões de movimentos para os objetos de cada um dos grupos. Essas trajetórias são utilizadas para determinar um padrão de movimentação e gerar um modelo para cada categoria. É utilizado o Modelo de Misturas Gaussianas (GMM, acrônimo de *Gaussian Mixture Models*, em inglês) para determinar os padrões primários de movimento. A abordagem proposta demonstrou melhor desempenho, comparada às abordagens de classificação baseadas no *Adaboost* e no *Latent Dirichlet Allocation (LDA)*.

Han et al. [2009] apresentam um descritor contextual e um método de aprendizado bayesiano para realizar a tarefa de reconhecimento de ações humanas. É proposto o uso do *bag-of-detectors*, que combina descritores espaço-temporais, Histograma de Gradientes (HoG), Histograma de Fluxo (HoF) e um gradiente 3D baseado no HoG, para descrição do conteúdo visual. A abordagem de classificação é construída a partir de um Classificador Bayesiano baseado em Processos Gaussianos com MultiKernel (MKGPC), aplicado nas bases de dados do KHT e Hollywood-1, registrando um dos melhores resultados já alcançados em ambos os conjuntos de dados.

Paleari et al. [2009] utilizam características multimodais, extraídas de informações audiovisuais, para a realização da tarefa de reconhecimento de emoções em cenas de vídeos digitais. É proposta a construção de uma rede neural baseada em modelos gerados por classificadores SVM e *nearest neighbors* (vizinho mais próximo). A abordagem foi comparada com diferentes métodos para fusão de evidências, entre eles o Bayesiano e a Votação, e demonstrou bons resultados na detecção de algumas das categorias.

O classificador *Naïve Bayes* (Bayes Ingênuo) é associado ao Modelo Oculto de Markov (HMM, acrônimo de *Hidden Markov Models*) em uma abordagem de classificação de gênero de vídeos proposta por You et al. [2010]. Foi observado que existe uma relação contextual entre os eventos que ocorrem na cena e o gênero dos vídeos. São extraídas características visuais globais e locais em conjunto com informações de movimento de câmera, padrão dominante de cor, entre outras, para caracterizar os eventos presentes nas cenas. A abordagem demonstrou bons resultados, tendo utilizado um conjunto pré-definido de eventos como parâmetro para ajuste do modelo.

Zhang e Gong [2010] apresentam uma abordagem para identificação de ações baseada no aprendizado da distribuição de tópicos latentes. É considerado que os quadros contendo as ações são obtidos sem interferência de movimentos de câmera e presença de plano-de-fundo. É realizada a descrição da silhueta da cena para representar a informação visual e construído um modelo hierárquico, baseado no *pLSA* (*Probabilistic Latent Semantic Analysis*). Os resultados obtidos foram superiores às abordagens tradicionais, tais como SVM, HMM e GMM.

A transferência de aprendizado é utilizada por Lopes et al. [2011], adotando uma base auxiliar de imagens para representar os conceitos, cada um dos quais possui um modelo

específico que considera também seus exemplos negativos. Para cada *frame* é estimada a probabilidade de pertencer a cada um dos conceitos e, posteriormente, essas probabilidades são combinadas dando origem a um vetor de probabilidades para o vídeo. O conceito que representa o vídeo é definido pela combinação desse vetor com uma probabilidade estimada por um classificador-base. Os experimentos foram conduzidos na base Caltech256 e Hollywood2, utilizando o classificador SVM como classificador-base.

Uma extensão do descritor espaço-temporal STIP é apresentada por Souza et al. [2011], denominada *HueSTIP*, que permite a incorporação da informação de cor. Os experimentos na aplicação de reconhecimento de ações demonstram resultados superiores à versão original, para as tarefas de detecção de ações na base Hollywood2.

Uma revisão de vários métodos propostos para a detecção de ações e eventos em vídeos é apresentada por Ballan et al. [2010]. E Poppe [2010] realiza um estudo de diferentes abordagens para a detecção de ações humanas em vídeos. Já Souza Jr et al. [2011] apresentam um resumo das técnicas de detecção automática de eventos em vídeos de jogos de futebol e são discutidas informações relevantes acerca das estratégias que podem ser adotadas em vídeos nesse domínio.

3.4 Classificação de vídeos e imagens em categorias semânticas de âmbito geral

Os trabalhos listados nesse subgrupo se concentram na categorização de vídeos em classes gerais, de acordo com suas características semânticas. Tarefas típicas envolvem determinar a presença de uma classe de objetos (por exemplo, pessoas, carros, etc.), decidir o gênero do vídeo (clipes musicais, telejornal, jogos de futebol, etc.) ou até mesmo prever se o usuário vai gostar daquele vídeo (recomendação). Novamente, a dependência a um contexto composto por outros vídeos, essencial para o presente trabalho, tem sido pouco explorada.

O *framework* proposto por Spyrou et. al. [2009] é utilizado na identificação de informações semânticas em sequências de vídeos analisando características visuais de cor e textura. É construído um *thesaurus* para relacionar um conjunto de vetores de características a uma descrição semântica. É utilizada a Análise de Semântica Latente (LSA,

acrônimo de *Latent Semantic Analysis*) para explorar as relações latentes no vetor de características; e um classificador SVM para treinar os conceitos. Os experimentos foram realizados na base de vídeos do TRECVID 2007.

Guillaumin et al. [2010] apresentam uma estratégia multimodal para a classificação de imagens. O conjunto de treinamento é constituído por duas partes: a primeira, contendo imagens associadas a um conjunto de palavras-chave e uma classe, e; a segunda, contendo imagens associadas apenas ao conjunto de palavras-chave. É utilizado um classificador *multikernel*, analisando as características visuais e textuais do primeiro conjunto, para gerar uma pontuação para as imagens do segundo grupo. O conhecimento é incorporado ao conjunto de treinamento e empregado na classificação de imagens desconhecidas, com base apenas na informação visual. Os experimentos foram realizados na base de dados do PASCAL VOC 2007, utilizando o classificador SVM.

Uma proposta para a classificação não supervisionada de imagens na web utilizando o *pLSA* é apresentada por Fergus et al. [2010]. São combinadas descrições textuais e visuais, no intuito de descrever o conteúdo das imagens e fornecer uma localização contextual para os objetos. Os experimentos conduzidos no *re-ranking* de buscas por imagens do *Google Image* e na classificação semântica utilizando a base de dados da CALTECH e PASCAL VOC 2007, demonstraram resultados satisfatórios.

Wu et al. [2011] apresentam uma alternativa para adicionar informação contextual à classificação de vídeos em redes sociais na web onde é utilizada a fusão de informação de interesse do usuário com a semântica e relevância do vídeo. A informação semântica é extraída de evidências textuais, fornecidas por metadados que acompanham o vídeo. A relevância de um vídeo é estimada de acordo com os vídeos que aparecem como relacionados a ele. O interesse do usuário é induzido a partir da análise dos demais vídeos postados pelo mesmo usuário. Cada evidência utiliza um classificador SVM independente, fornecendo uma probabilidade da distribuição do vídeo nas diferentes categorias possíveis (música, comédia, entretenimento, entre outras). A classificação do vídeo é baseada no somatório dessas probabilidades. Os experimentos foram executados em um conjunto de vídeos obtidos junto ao YouTube, apresentando resultados satisfatórios.

Song et al. [2011] propõem uma abordagem para adicionar contexto à classificação de imagens, denominada *Context-SVM*. São empregadas etapas de classificação e detecção

de objetos. Inicialmente é realizada a classificação das imagens, gerando uma estimativa de probabilidade de a imagem pertencer a cada uma das possíveis categorias. Esse resultado é utilizado como indicativo de contexto para uma etapa de detecção de objetos. O resultado da detecção é utilizado como indicativo de contexto para uma nova etapa de classificação de imagens. Os experimentos foram conduzidos na base de dados do PASCAL VOC 2007 e 2010, obtendo os melhores resultados de classificação para a maioria das categorias nesse último.

3.5 Resumo estático de vídeos

O processo de geração de resumos automáticos é essencial para uma série de aplicações em vídeos digitais, como indexação, navegação, categorização e recuperação por conteúdo [Truong e Venkatesh, 2007]. Li et al. [2001] apresentam divisão das técnicas para a geração de resumos automáticos que podem ser divididas em duas grandes categorias: resumos estáticos, conhecida como *keyframes* ou *video-skim* (em inglês), e; resumos dinâmicos, conhecida como *dynamic summary* (em inglês). Aqui apresentamos apenas a abordagem de resumos estáticos, devido a sua utilização no nosso trabalho. Um trabalho recente para resumos dinâmicos foi elaborado por Cunha et al. [2012] onde os experimentos com vídeos não editados (*rushes videos*, em inglês) demonstraram resultados comparáveis com as melhores estratégias do TRECVID 2007.

Os resumos estáticos são a representação do conteúdo visual do vídeo através de um conjunto de quadros-chave capazes de sintetizar a informação visual contida no vídeo. Na literatura da área são encontradas várias técnicas para a realização da tarefa de geração automática de resumos estáticos, tais como [Zhuang et al., 1998; Hadi et al., 2006; Mundur et al., 2006; Chang e Chen, 2007; Furini et al., 2007]. É característica comum a todas essas abordagens a existência de pelo menos duas etapas: “segmentação temporal do vídeo” e “seleção de quadros-chave”. Após a realização dessas etapas comuns, as diferentes abordagens executam tarefas de refinamento e organização da informação a ser apresentada. Cotsaces et al. [2006] e Yuan et al. [2007] apresentam apanhados gerais sobre as técnicas para segmentação temporal de vídeo. Já Truong e Venkatesh [2007] apresentam uma revisão sistemática das técnicas para a identificação de quadros-chave e

as subdividem em diferentes categorias, de acordo com suas características: mudança significativa de conteúdo, variância temporal semelhante, cobertura máxima do quadro, agrupamento, correlação mínima entre quadros-chave, erro na reconstrução da sequência, simplificação de curvas e eventos interessantes.

Neste trabalho foi adotado o método *VSUMM* (*Video SUMM*ary em inglês - Resumo de Vídeo) apresentado por Avila et al. [2008a]. Essa abordagem demonstrou melhores resultados que outras apresentadas na literatura, conforme observado nos experimentos descritos por Avila et al. [2008a, 2008b]¹¹. O método foi desenvolvido para ser simples e eficiente, utilizando atributos globais de baixo-nível para descrever o conteúdo visual presente nos vídeos e realizando uma análise direta dos quadros que compõem o vídeo, não sendo necessário fazer a segmentação temporal.

Outra vantagem do *VSUMM* é a geração dos agrupamentos de quadros similares independentemente da tomada a que cada quadro pertence. Com isso, caso algum quadro se repita no decorrer do vídeo, eles são identificados e pode-se evitar que ambos apareçam no resumo gerado.

3.6 Extratores de características visuais

Para realizar a classificação de imagens baseada na similaridade de conteúdo visual é necessário utilizar ferramentas que realizem a descrição de características que representem esse conteúdo. Aqui, discutimos as chamadas características visuais de baixo-nível, que são extraídas diretamente dos *pixels* e *frames* e constituem a primeira etapa essencial na obtenção de uma representação útil para busca e classificação de documentos visuais. Na literatura contemporânea, essa etapa é frequentemente seguida pela criação de uma representação de nível-médio, construída pela agregação de características de baixo-nível. Uma das formas mais populares de obter essa

¹¹ Por ser uma técnica utilizada em etapa de pré-processamento de vídeos, anterior a extração de características, foi adotada a referência válida para a época de início do trabalho. Atualmente podem ser encontradas novas publicações com resultados mais promissores como é o caso do trabalho elaborado por Almeida et al. [2012].

representação de nível-médio é o modelo de *bags-of-visual-features*, que será explicado na próxima seção.

Torres e Falcão [2006] observam uma “confusão” na literatura com relação aos termos “descriptor” e “extrator” de características visuais, sendo que os descritores devem ser compostos por duas partes: a ferramenta de extração de características e uma função de distância capaz de medir a similaridade entre um par de características.

A literatura apresenta diferentes abordagens para extração de características que podem ser utilizadas para representação do conteúdo visual de imagens, como: cor, textura e forma dos objetos, etc. Essas características são organizadas em um vetor multidimensional que é utilizado como assinatura do conteúdo visual presente. Esse vetor é conhecido como “vetor de características” [Gonzalez e Woods, 2006].

Considerando uma interpretação em mais baixo-nível, um vídeo pode ser entendido como um conjunto de imagens. Por isso, além de técnicas específicas para vídeos, podem ser utilizadas técnicas de imagens. Contudo, para a aplicação efetiva dessas técnicas, faz-se necessário um passo adicional de pré-processamento, tal como o visto na seção anterior, para reduzir a redundância das informações extraídas.

3.6.1 Extratores de características globais

Os extratores globais consideram o conteúdo integral da imagem ou vídeo para análise das características visuais, não necessitando de particionamento em regiões ou identificação de pontos de interesse. Geralmente, essas técnicas exigem baixo custo computacional para execução, o que pode ser uma grande vantagem de acordo com a aplicação. Contudo, por considerar a imagem como um todo em sua análise, são pouco robustas as alterações de plano-de-fundo e oclusões.

Neste trabalho são adotadas as técnicas de histograma de cor nos canais RGB, histograma de matiz e momentos de Zernike, como representantes das abordagens globais. As duas primeiras estão relacionadas à distribuição de cores presentes na imagem. A terceira técnica está relacionada à representação global da forma.

O histograma de cor nos canais RGB foi adotado por ser uma das técnicas mais tradicionais na representação de conteúdo visual. O uso do espaço HSV para a construção do histograma de cor também demonstraram resultados interessantes como os

apresentados por Ma e Zhang [1998], onde foi apontado como o melhor dentre todos os espaços para construção de histogramas para representação de conteúdo em imagens. A simplificação, adotando apenas o canal *H* (*Hue*, Matiz), relacionado à homogeneidade das cores, se deve ao bom resultado obtido em experimentos preliminares. Já Li et al. [2009] e Goyal e Walia [2012] observam a importância do uso de momentos de Zernike como uma estratégia global para a descrição de forma em imagens, motivando a adoção dessa técnica neste trabalho.

Kimura et al. [2011] e Penatti et al. [2012] apresentam *surveys* comparativos entre diversas técnicas de extração de características visuais que atuam no âmbito global.

Histograma de cor nos canais RGB

Um histograma descreve a distribuição de uma medida de interesse, observando as frequências dos valores dessa medida em uma população ou amostra. Assim, o histograma de cor de uma imagem descreve a frequência relativa das cores encontradas nessa imagem [Gonzalez e Woods, 2006]. O método é de fácil compreensão intuitiva e computacionalmente simples e, por isso, bastante utilizado em aplicações de recuperação de imagens por conteúdo.

O histograma de cor mais tradicional utiliza uma quantização do espaço de cor como um todo para prover um único histograma que caracteriza a distribuição das cores das imagens. Fica claro que a qualidade da representação dependerá tanto da escolha do espaço de cor (RGB, HSV, Lab, etc.) quanto da forma como este é quantizado (regular, adaptativa aos dados, etc.) para a confecção do histograma.

Neste trabalho, entretanto, utilizamos a versão do histograma de cor que caracteriza, separada e independentemente, cada um dos canais de cor do espaço RGB, estabelecendo um histograma de valores para cada canal. O número de dimensões do vetor de características gerado por este atributo é igual ao número de cores possíveis de ocorrerem na imagem, multiplicado pelo número de canais de cor do espaço utilizado, 768 dimensões (256 cores * 3 canais).

Histograma de matiz

O matiz, ou tonalidade, é a sensação visual de tonalidade das cores: vermelha, amarela, verde e azul, ou uma combinação dessas. A informação de matiz verifica a cor pura, cor sem a adição de branco ou preto, observada no *pixel*, abrangendo todas as cores do espectro.

O histograma é construído de modo similar ao histograma de cor, porém, o espaço de cores utilizado é o HSV, sendo usado apenas no canal *H*. O número de dimensões do vetor de características gerado por este atributo é igual ao número de cores presentes na dimensão *H* do cone *HSV*, usualmente, 360 dimensões.

Momentos de Zernike

Os momentos são estatísticas que estabelecem o formato geral de uma distribuição. Considerando uma imagem como uma distribuição de valores de intensidade ao longo do seu domínio espacial (x, y) , os momentos permitem obter uma caracterização compacta e robusta da forma global da imagem. Os coeficientes de momento podem ser combinados para obter diversas invariâncias (translação, escala, rotação, etc.), embora a cada invariância desejada, as equações vão ficando cada vez mais complexas.

Os polinômios complexos propostos por Zernike [1934] combinam os momentos de modo a capturar a forma global da imagem ao mesmo tempo em que preservam algumas informações locais, conforme observado por Ronda et. al. [2008], sendo eficazes na representação de formas complexas e compostas por várias regiões disjuntas.

A proposta original é invariante à rotação, mas Khotanzad e Hong [1990] propuseram uma etapa adicional de normalização para possibilitar a invariância à escala e translação. Na presente tese será adotada essa versão adaptada.

3.6.2 Extratores de características locais

Schmid e Mohr [1996] foram um dos primeiros a utilizarem extratores locais para representar o conteúdo visual em tarefas de reconhecimento de objetos. A ideia é que essas técnicas sejam invariantes às transformações geométricas (mudanças de escala e rotação) e fotométricas (alteração no padrão de cores e iluminação). O objetivo é a

detecção de pontos, ou regiões, de interesse que sejam capazes de representar o conteúdo da imagem.

Esta subseção apresenta algumas abordagens para a extração de atributos visuais locais, descrevendo apenas aqueles que foram utilizados neste trabalho. Mikolajczyk et al. [2005] e Tuytelaars e Mikolajczyk [2008] apresentam um *survey* comparativo de técnicas para detecção de regiões de interesse em imagens. Mikolajczyk e Schmid [2005] e Zhang et al. [2007] apresentam *surveys* comparativos de técnicas para extração de características locais. Esses trabalhos são as principais referências exaustivas do tema de extração de regiões de interesse e características locais em imagens.

Este trabalho avalia o desempenho de três extratores locais puramente espaciais (SIFT, PCA-SIFT e HUE-SIFT) e um espaço-temporal (STIP). O SIFT e PCA-SIFT foram selecionados devido ao bom desempenho apresentado no estudo comparativo de Mikolajczyk e Schmid [2005], além do fato de o SIFT ser o extrator local mais utilizado na literatura recente. O HUE-SIFT foi selecionado por ter demonstrado um desempenho similar ou superior ao SIFT nos resultados da classificação de vídeos, apresentado por Van de Sande et al. [2010]. E o STIP, além de ser um dos extratores espaço-temporal mais utilizados na literatura, apresentou bons resultados nos experimentos realizados por Lopes et al. [2009c] e Souza et al. [2010], ambos relacionados à classificação semântica de vídeos.

SIFT

Proposto por Lowe [1999], o SIFT, *Scale-Invariant Feature Transform* (Transformada de Característica Invariante à Escala), é um método invariante à escala e à rotação, além de possuir bom desempenho a alterações de iluminação, ruído, pequenas mudanças no ponto-de-vista e oclusões [Mikolajczyk e Schmid, 2005].

Na proposta original é sugerida a utilização da verificação de pontos de máximo e mínimo para a função de diferenças de gaussianas - *DoG* (*Difference-of-Gaussian*, em inglês) como abordagem para detecção das regiões de interesse. A técnica de detecção em diferenças de gaussianas estabelece uma escala-espaço em que a escala da detecção permite obter invariância a escala. Uma etapa de estimativa de direção principal do gradiente permite obter invariância à rotação. Uma vez que a região em torno do ponto de interesse é colocada em coordenadas invariantes, o descritor pode ser obtido através

de uma concatenação de histogramas de gradientes em pequenas subjanelas (normalmente 16, numa grade 4x4) da região de interesse. A dimensionalidade do vetor de características resultante é 128.

Existem na literatura muitas variações do SIFT. Duas delas, utilizadas neste trabalho, são:

- **HUE-SIFT:** o HUE-SIFT [Van de Weijer et al., 2006] foi desenvolvido como sendo uma extensão que, além da informação de gradiente e orientação extraída pelo SIFT, analisa a informação de cor, considerando o canal *H*. Com isso, este extrator, além de invariante às mudanças de escala e rotação, torna-se resistente às mudanças de intensidade de cor. O vetor de características resultante possui 165 dimensões;
- **PCA-SIFT:** no PCA-SIFT [Ke e Sukthankar, 2004] é feita a concatenação do gradiente horizontal e vertical utilizando uma matriz 39x39, centralizado e rotacionado em um ponto de interesse para alinhar a orientação dominante com uma direção padrão. É produzido um vetor de características com 3042 dimensões, no qual é aplicada a PCA [Pearson, 1901], *Principal Components Analysis* (Análise de Componentes Principais), para promover a redução de dimensionalidade. São selecionados apenas os *n* maiores autovetores (*eigenvectors*, em inglês) resultantes da aplicação do PCA para comporem o vetor de características. Nos experimentos realizados por Ke e Sukthankar [2004] o *n=36* apresentou melhor desempenho. Com isso, o vetor de características produzido possui 36 dimensões.

STIP

O *Space-Time Interest Point* – STIP [Laptev, 2005] - Ponto de Interesse Espaço-Tempo - é apropriado para representar o conteúdo visual em vídeos. No trabalho original é apresentada uma extensão ao operador de Harris tornando-o capaz de analisar tanto a dimensão espacial quanto temporal, no intuito de servir como detector de regiões de interesse no espaço-tempo. São propostos três extratores de características visuais, *HOG* (*Histograms of Oriented Gradients*, Histogramas de Gradientes Orientados) e *HOF* (*Histograms of Optical Flow*, Histogramas de Fluxo Óptico); o HOG-HOF, uma concatenação desses anteriores. Enquanto o HOG foca em características da aparência do

conteúdo, o HOF descreve aspectos do movimento. É gerado um vetor de características com 162 dimensões, HOG (72) e HOF (90), invariante às transformações geométricas, mudanças de iluminação ou sombreamento. Quando utilizado neste trabalho é adotado o extrator HOG-HOF.

3.7 Representação de características em nível-médio

A representação de conteúdo visual gerada pelos extratores de características, tal como os apresentados na Seção 3.6, frequentemente permite discriminar precisamente objetos específicos, especialmente aqueles obtidos pelos descritores locais. Entretanto, essas representações não propiciam a capacidade de generalização necessária à tarefa de classificação em categorias semânticas.

Uma alternativa é o uso do *bag-of-visual-features* (bolsa de características visuais), baseada na técnica de *bag-of-words* (bolsa de palavras) [Sivic e Zisserman, 2003], já consolidada na área de recuperação de informação textual. A abordagem de *bag-of-visual-features* adiciona um nível de significância às características ao utilizá-las de forma agregada. Além disso, a representação é mais conveniente, pois se obtém um vetor de características de tamanho fixo, ao invés de um conjunto de tamanho variável de características locais.

O uso desse modelo tem se tornado um importante tópico nas atividades de pesquisa na área de visão computacional, a comprovação disso é o grande volume de abordagens utilizando *bag-of-visual-features* encontradas na literatura recente da área, como visto nos trabalhos de Yang et al. [2007b], Jiang et al. [2007], Laptev et al. [2008], Niebles et al. [2006], Liu et al. [2008], Deselaers et al. [2008], Lopes et al. [2009a], Kinnunnen et al. [2009], Boureau et al. [2010], Jégou et al. [2010] e Ding et al. [2011]. Essa técnica também é conhecida por outros nomes como: *textons* [Leung e Malik, 2001], *object parts* [Fergus et al., 2003] e *codebooks* [Jurie e Triggs, 2005].

Uma descrição do uso de *bag-of-visual-features* no contexto de detecção de objetos em imagens é apresentada por Sivic e Zisserman [2008] e serve como uma boa referência para a compreensão geral dessa técnica. A Figura 3.1 apresenta uma estrutura genérica da

abordagem de *bag-of-visual-features* que é executada em duas etapas principais, após uma etapa de descrição do conteúdo visual:

- Inicialmente é construído o vocabulário visual. Existem diversas estratégias na literatura para a realização dessa etapa, Tian et al. [2011] apresentam um estudo dessas abordagens. Comumente são utilizadas técnicas de agrupamento, tal como o *k-means*, para gerar os vocabulários. Nessa situação os centróides de cada agrupamento são considerados como sendo uma palavra visual e o conjunto dessas palavras formam o vocabulário. Neste trabalho foi utilizada uma estratégia de seleção aleatória das palavras do vocabulário, com base no conjunto de pontos-de-interesses identificados e descritos em todos os vídeos. Essa abordagem foi adotada com base na observação de experimentos preliminares que demonstraram igual desempenho exigindo menor custo de processamento se comparado ao uso de agrupamentos;
- Na etapa de quantização, todas as imagens ou vídeos são avaliados em relação ao vocabulário visual, gerando um histograma onde cada dimensão está associada a uma palavra do vocabulário. É medida a similaridade entre cada um dos pontos-de-interesse na imagem ou vídeo (entendidas como sendo as palavras visuais presentes no objeto avaliado) e cada uma das palavras do vocabulário, a dimensão que representa a palavra visual com maior similaridade tem seu valor incrementado, indicando a presença daquela palavra no objeto avaliado. Diferentes normalizações podem ser aplicadas (TF, IDF, TF-IDF, etc.) O histograma resultante dessa etapa é conhecido como *bag-of-visual-features* por representar em um mesmo conjunto todas as palavras visuais sem nenhum tipo de relação espacial, hierárquica ou de dependência entre elas.

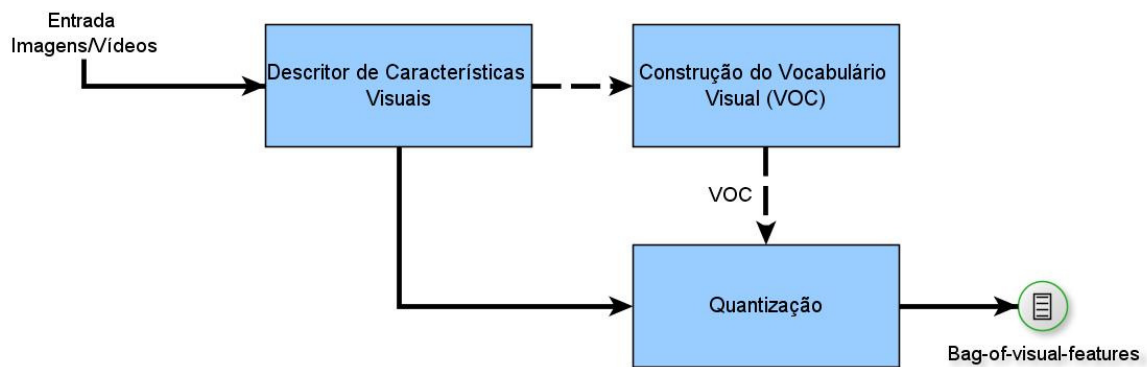


Figura 3.1. Estrutura genérica da abordagem de *bag-of-visual-features*.

A abordagem de *bag-of-visual-features* tem demonstrado bons resultados em diferentes tarefas de identificação de objetos e cenas. Contudo a simplicidade da representação é ao mesmo tempo seu ponto forte e seu ponto-fraco, pois informações de dependência ou espacialidade das palavras visuais são desprezadas. Mais essencialmente, no tocante ao presente trabalho, a representação não incorpora nenhuma informação de contexto do vídeo, sendo necessário tratar esse contexto à parte. Esse último aspecto levará à principal contribuição desta tese, que é a possibilidade de incorporar a informação de contexto diretamente no modelo de *bags*, conforme descreve-se nos Capítulos 6 e 7.

3.8 Classificação supervisionada

Classificador é uma ferramenta computacional capaz de identificar a categoria de um elemento. Um classificador é supervisionado se depende da existência de um conjunto de dados previamente classificado, utilizado como referência para a categorização dos demais; caso contrário é não supervisionado. Neste trabalho são abordados apenas classificadores supervisionados.

De modo genérico, a tarefa de classificação supervisionada consiste em realizar a classificação de novos elementos, com base em conjunto de dados anotado.

Uma técnica de classificação supervisionada pode ser baseada em uma dentre duas abordagens: discriminativa ou generativa. Os classificadores discriminativos constroem um modelo de mapeamento dos dados de entrada, x , para uma classe, y . São representantes desse conjunto o SVM – *Support Vector Machines*, Máquinas de Vetores

de Suporte [Vapnik, 1979], Boosting [Mason et al., 2000], Redes Neurais [Arbib, 1995], Regressão Logística [Hosmer e Lemeshow, 2000], entre outros.

Os classificadores generativos aprendem um modelo com base na probabilidade conjunta $p(x, y)$, onde x é a entrada e y a classe, e fazem uma classificação estimativa com base na regra de Bayes, calculando $p(y | x)$. São representantes desse modelo: *GMM* – *Gaussian Mixture Model*, Modelo de Misturas Gausiano [Permuter et al., 2003]; *HMM* – *Hidden Markov Model*, Modelo Oculto de Markov [Rabiner, 1989]; *Naïve Bayes*, Bayes Ingênuo [Ng e Jordan, 2001]; *LDA* – *Latent Dirichlet Allocation*, Alocação Latente de Dirichlet [Blei et al., 2003]; *PLSA* – *Probabilistic Latent Semantic Analysis*, Análise Probabilística de Semântica Latente [Hofmann, 1999], entre outras.

Neste trabalho são avaliados os classificadores SVM e *Naïve Bayes*, técnicas tradicionais na literatura.

3.8.1 SVM

O SVM - *Support Vector Machines*, Vetor de Máquinas de Suporte, é uma técnica de classificação proposta por Vapnik [1979]. O SVM particiona os dados em um espaço euclidiano traçando hiperplanos de modo a isolar os elementos de uma mesma classe em cada uma das regiões geradas. Um tutorial sobre o uso do SVM é apresentado por Burges [1998].

A Figura 3.2, apresenta um exemplo de identificação de hiperplanos para a separação de elementos de duas classes. São identificados três hiperplanos possíveis e é selecionado pelo SVM aquele que está há maior distância dos elementos de ambas as classes. Os vetores de características que são mais próximos do hiperplano são chamados de *support vectors* (vetores de suporte, em português), significando que a posição dos outros vetores não afeta o hiperplano. No exemplo apresentado o hiperplano *h2* representa a melhor opção, por induzir a maior distância entre o hiperplano e os vetores de suporte.

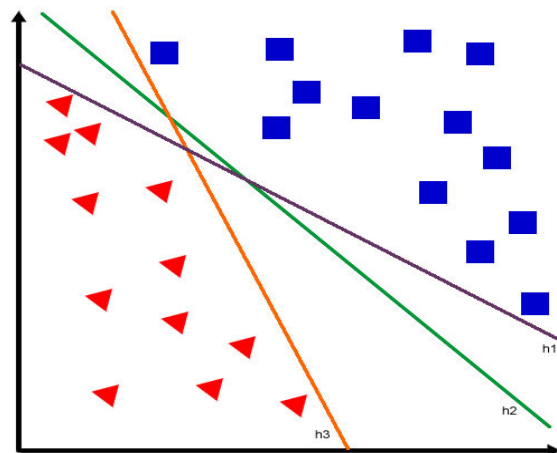


Figura 3.2. Processo de identificação de hiperplano para separação de conjunto de dados. O hiperplano $h2$ representa a melhor opção, por obter a maior distância entre o hiperplano e os vetores de suporte.

3.8.2 Naïve bayes

O classificador *naïve bayes* é um classificador probabilístico e fundamentado na aplicação do teorema de Bayes. A característica mais marcante deste classificador, responsável pela atribuição do nome *naïve* (ingênuo), é o fato de considerar a independência total entre todos os atributos. Ou seja, no seu funcionamento o classificador não considera nenhum tipo de correlação existente entre os atributos, ou variáveis, que compõem o vetor de características.

A hipótese de independência de variáveis parece, em um primeiro momento, ser extremamente forte, suscitando a desconfiança desse classificador jamais obter resultados satisfatórios. Contudo, contrariando essa expectativa, Caruana e Niculescu-Mizil [2006] demonstram que o *naïve bayes* supera vários outros classificadores de abordagem supervisionada, em muitas situações. Uma revisão do uso do *naïve bayes* é apresentada por Zhang [2004], discutindo situações em que pode ser adotado.

O teorema de Bayes tem diferentes interpretações, mas na inferência bayesiana, ele está relacionado com a mudança em nossa crença após efetuar uma observação. A probabilidade condicional (ou *a posteriori*) da nossa crença X após efetuar a observação C é dada pela probabilidade marginal (ou *a priori*) da crença X antes da observação, multiplicada por quanto a crença X torna provável a observação C , dada pelo chamado termo de verossimilhança $p(C | X)$, e normalizada pela probabilidade da observação:

$$p(X | C) = \frac{p(X)p(C | X)}{p(C)},$$

Dessa forma, o teorema de Bayes nos permite atualizar nossas crenças (por exemplo, do grau de pertinência de uma imagem à classe *spam* ou à classe legítimo) em face de novas evidências observadas (por exemplo, o vetor de características obtido).

3.9 Combinação de evidências

Segundo Snoek et al. [2005], as abordagens de fusão de evidências podem ser divididas em dois grupos: a) *a priori*, ou *early fusion*, e; b) *a posteriori*, ou *late fusion*, de acordo com o instante em que são adotadas no processo de classificação.

As abordagens *a priori* são caracterizadas por unir as informações extraídas de diferentes origens em um único elemento composto. Ou seja, primeiramente as informações são extraídas e representadas de modo isolado e, então, é gerada uma nova representação, combinando todas as previamente obtidas. Com esse modelo, uma vantagem é a dependência de apenas um classificador, enquanto que a dificuldade é elaboração de uma representação capaz de absorver as informações oriundas das diferentes fontes.

Na abordagem *a posteriori*, a combinação de evidências é realizada após a obtenção do resultado da classificação com base nas evidências isoladas. É adotada uma estratégia para se decidir a qual categoria cada elemento analisado deve ser inserido, de acordo com a análise conjunta dos resultados individuais. Esse tipo de técnica demanda maior custo computacional para a obtenção das evidências, visto que cada elemento está associado a um classificador individual, implicando em etapas de treinamento separadas. Além disso, após a obtenção das evidências isoladas pode ser necessário uma etapa de treinamento adicional para realizar a classificação com as evidências combinadas. Contudo, essas técnicas apresentaram melhores resultados que as anteriores na classificação semântica de vídeos, conforme observado por Snoek et. al. [2005].

Segundo verificado por Benmokhtar e Huet [2006], as abordagens para a execução da combinação de evidências *a posteriori* podem ser divididas em três categorias: baseadas em meta-classificadores, supervisionadas e não supervisionadas.

As abordagens baseadas em meta-classificadores são caracterizadas por utilizarem versões modificadas do conjunto de dados de treinamento. Os resultados de cada classificador são então combinados. Algumas abordagens nessa categoria são o *Adaboost* [Freud e Schapire, 1996], *Bagging* [Skurichina e Duin, 1998] e *Random Subspace* [Skurichina e Duin, 1998], que consiste no uso conjunto de *Bagging* e *Boosting*.

As abordagens do tipo supervisionadas são técnicas que necessitam de uma etapa prévia de aprendizado para realizar o processo de combinação de evidências. Segundo Benmokhtar e Huet [2006], as principais são as que se enquadram nos seguintes grupos: Redes Neurais, Modelos de Mistura Gaussianas, Modelos de Decisão e Algoritmos Genéticos.

Nas abordagens não supervisionadas, se enquadram as técnicas de combinação de evidências que não necessitam de um treinamento prévio para realizarem a fusão das informações multimodais, geralmente usando votação. Uma das variações é a conhecida como “maioria absoluta” (*absolute majority*, em inglês), e que um elemento é classificado de acordo com a opinião de mais da metade dos classificadores individuais. Já na “maioria simples” (*simple majority*, em inglês) a classificação é de acordo com a categoria que obtiver a maior quantidade dos votos. Outra variação possível é a adoção de uma ponderação para cada classificador, definindo a importância de seu voto na classificação final, denominada de “mais votado ponderado” (*weighed majority voting*, em inglês).

Este trabalho adota uma estratégia de combinação de evidências baseada em votação com maioria simples/maioria absoluta (os dois se equivalem para problemas com duas classes). Fora realizadas tentativas preliminares de ponderar os votos pelo grau de confiança da classificação, mas não se observou melhorias.

3.10 Adequação do espaço geométrico

Muitas aplicações recentes de aprendizagem de máquina necessitam explorar um conjunto de dados com um grande número de exemplos, cada um dos quais é composto de muitos atributos (ou dimensões). Essa situação acaba por dificultar a tarefa na maioria dos algoritmos tradicionais. Uma alternativa para auxiliar nesse problema é a projeção dos dados em outro espaço geométrico mais conveniente, por ter dimensionalidade reduzida,

distâncias mais homogêneas, maior separação entre as classes, etc. A transformação de espaços pode ser **Linear**, como na Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) e Análise de Semântica Latente ou **Não-Linear** (também conhecido como *Manifold Learning*, Aprendizado em Variedades), como: *Locally Linear Embedding* (LLE) [Roweis e Lawrence, 2000], *Laplacian Eigenmaps* [Belkin e Niyogi, 2001] e *Isomaps* (Tenenbaum et al., 2000). Em Pless e Souvenir [2009] é apresentado um estudo de técnicas de *manifold learning* aplicadas a imagens.

Neste trabalho são avaliadas apenas técnicas lineares, mais especificamente a Análise de Semântica Latente. O LSA foi adotado por ter demonstrado resultados promissores em aplicações similares, além de ser a mais simples e tradicional das abordagens de análise de semântica latente.

3.10.1 Análise de Semântica Latente

A adoção de técnicas para explorar as correlações nas características tem se tornado comum na literatura. Para realizar esse processo são utilizadas abordagens de Análise de Semântica Latente. Dentre essas se destacam a LSA (*Latent Semantic Analysis*, Análise de Semântica Latente), pLSA (*Probabilistic Latent Semantic Analysis*, Análise Probabilística de Semântica Latente) e LDA (*Latent Dirichlet Allocation*, Alocação Latente de Dirichlet).

Essas técnicas foram inicialmente adotadas para realizar a identificação da semântica latente em vetor de características representando o modelo de *bag-of-words*, largamente utilizado na área de recuperação de informação textual. Contudo, estão sendo utilizadas com sucesso em aplicações envolvendo a descrição de conteúdo visual em imagens e vídeos, tal como demonstrado por diversos autores. Monay e Gatica-Perez [2007] avaliaram diferentes alternativas para classificação de imagens em categorias semânticas utilizando o pLSA combinando evidências textuais e visuais. Niebles et al. [2006] utilizaram pLSA e LDA para o reconhecimento de ações humanas em cenas de vídeo com base em descritores espaço-temporais, observando um desempenho mais satisfatório com o uso do pLSA. Zhou et al. [2011] aplicaram a LSA associada à construção de um grafo conectado de imagens e palavras visuais, geradas com base nos descritores SIFT e MSER, na classificação de imagens retornadas em buscas textuais no *Google Image* verificando

aumento significativo na qualidade dos resultados em comparação ao uso tradicional de *bags-of-visual-features*.

Na LSA é realizada uma etapa de preparação dos dados, no sentido de normalizar os valores do vetor de características, geralmente utilizando a abordagem de TF-IDF [Gerard e Buckley, 1988]. Com base na matriz documento-termo, executa a Decomposição de Valores Singulares (SVD, acrônimo de *Singular Value Decomposition*).

A LSA nada mais é do que uma operação de conversão de espaços, com o objetivo de explicitar as associações latentes entre os documentos. A SVD é a responsável por realizar a conversão para o novo espaço geométrico, que passa a ser conhecido como espaço de tópicos latentes. Pode ser obtida uma redução de dimensionalidade aplicando o descarte de dimensões com valores singulares de baixa magnitude, que indicam tópicos pouco informativos no espaço.

Capítulo 4

Abordagens tradicionais

Este trabalho se insere no contexto de classificação de vídeos em categorias semânticas visando à detecção de vídeos não colaborativos em redes sociais. Essa aplicação possui características distintas e interessantes tanto com relação à natureza da base de vídeos (grande quantidade, curta duração e grande diversidade visual extraclasses) quanto com relação às classes que se deseja obter (semântica extremamente complexa e dependente do contexto, além de um conjunto de treinamento escasso).

Para dar andamento à investigação e elaboração de uma abordagem eficaz de detecção desses vídeos, é atacada a questão da extração de características visuais de baixo nível, perguntando-se qual tipo de descritor visual seria mais adequado, se a combinação de diferentes descritores seria útil, e qual seria a melhor forma de fazer essa combinação. Para auxiliar na resposta a essas perguntas, foi construída uma “solução modelo”, usando métodos tradicionais da literatura, mas permitindo a variação controlada dos fatores de interesse, notadamente as características visuais (extraídas no início do processo), e a combinação das evidências (efetuada no fim do processo).

A solução modelo não constitui a proposta original de solução do problema de detecção de ações não colaborativas, mas sim a definição de uma arquitetura tradicional e configurável, que permita a avaliação rigorosa das hipóteses iniciais. A proposta original deste trabalho é abordada nos Capítulos 6 e 7.

A solução modelo é apresentada na Seção 4.1 e o projeto experimental utilizado em sua avaliação é justificado nas Seções 4.2 a 4.6. Os resultados obtidos em cada um dos experimentos são apresentados no Capítulo 5.

4.1 “Solução modelo” com combinação de evidências

A solução modelo tem por característica ser modular e configurável, devido ao fato de ter sido concebida para possibilitar a avaliação de algumas das hipóteses de pesquisa deste trabalho. A Seção 1.4, apresenta as hipóteses de pesquisa, em caráter geral. As hipóteses a serem avaliadas pela solução modelo são reapresentadas abaixo, sendo que a hipótese de escolha do descritor visual de baixo-nível (*Hipótese 1*) está dividida em três etapas:

- **Hipótese 1.1:** devido à grande variabilidade de conteúdo dos vídeos, o uso de descritores de características visuais de baixo-nível local apresenta resultados de melhor qualidade que descritores globais;
- **Hipótese 1.2:** a aplicação de características visuais espaço-temporais obtém maior qualidade de resultados se comparadas às características puramente espaciais, por fazerem uso das relações entre os quadros do vídeo;
- **Hipótese 1.3:** devido às características particulares das ações não colaborativas abordadas, a utilização conjunta de outras informações, além das puramente visuais, extraídas do vídeo, contribui para o aumento da qualidade do resultado da classificação;
- **Hipótese 2:** o classificador SVM, por ser discriminativo e por não considerar as dimensões da representação de nível-médio como independentes, tem desempenho superior ao do *naïve bayes*, que é generativo e assume independência;
- **Hipótese 3:** a utilização conjunta de diferentes evidências visuais contribui para o aumento da qualidade dos resultados da classificação, se comparada ao uso isolado dessas evidências.

Dadas essas hipóteses, pode ser observado que a solução modelo deve ser capaz de: a) usar diferentes descritores visuais, tanto globais quanto locais; b) dentre os descritores locais, usar tanto os puramente espaciais quanto os espaço-temporais; c) considerar a opinião de mais de um classificador, utilizando técnicas de combinação de evidências *late fusion*, durante a etapa de tomada de decisão; d) permitir o uso de diferentes

classificadores, SVM e *naïve bayes*, e; e) possibilitar a inclusão de outras fontes de informação que não as puramente visuais.

Para atender a essas exigências, a solução modelo foi elaborada da seguinte forma:

- Inicialmente, o vídeo é encaminhado para a etapa de Pré-Processamento, onde ocorre a preparação do vídeo para ser submetido às técnicas de descrição de conteúdo visual. O vídeo é amostrado com taxa de um quadro por segundo e são selecionados os quadros-chave, com a finalidade de representar as informações principais presentes no vídeo, utilizando uma estratégia de resumo automático;
- Na etapa de Extração de Características, cada descritor de informação utiliza o elemento de vídeo mais adequado para sua execução (estatística estrutural do vídeo, quadros-chave ou sequência de vídeo completa). As informações geradas pelos descritores são padronizadas e encaminhadas, em separado, aos classificadores;
- Cada classificador está associado a uma característica visual e atua de modo independente. Cada um propõe uma classe ao vídeo de acordo com as informações que possui;
- As categorias sugeridas pelos classificadores são repassadas para a etapa de Combinação de Evidências, onde é adotada a estratégia do “mais votado” para a definição da classe a ser atribuída ao vídeo.

Na Figura 4.1, é apresentada a arquitetura da solução modelo construída com base na técnica de combinação de evidências *late fusion*. As etapas 2, 3 e 4 são objetos de observação no projeto experimental, que é apresentado nas próximas seções.

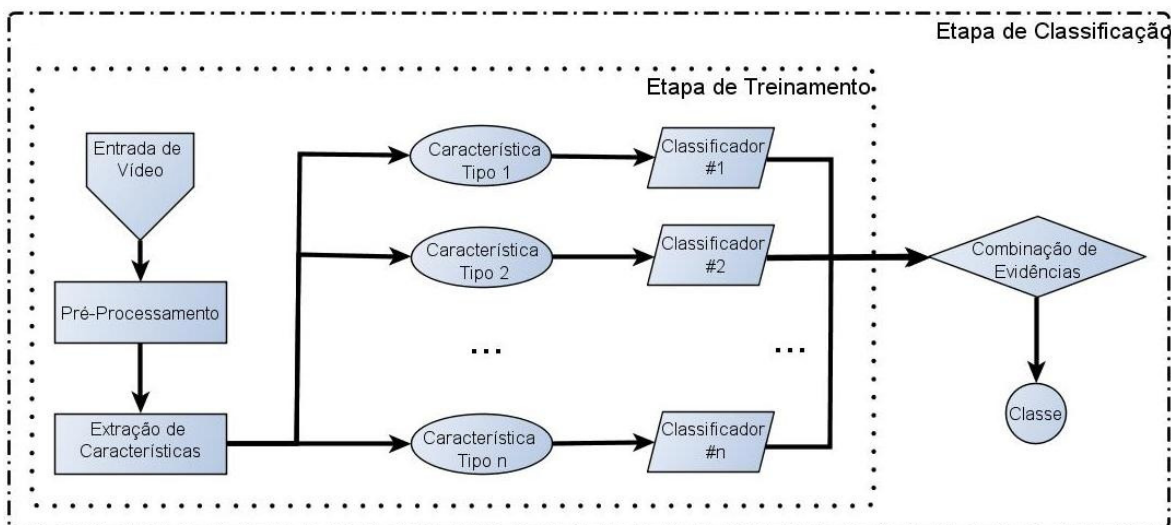


Figura 4.1. Arquitetura da “solução modelo” com combinação de evidências. As etapas de treinamento e classificação possuem execução similar, sendo que, na etapa de classificação é adicionada a etapa de combinação de evidências.

A etapa de Treinamento tem por objetivo gerar os modelos a serem utilizados durante a etapa de Classificação. Esta etapa consiste na execução das seguintes partes da arquitetura: a) Pré-processamento; b) Extração de Características, e; c) Construção dos Classificadores (Modelos). Na Figura 4.1, esta etapa é representada por uma linha pontilhada que circunda todas as partes que a integram.

Na etapa de Classificação, cada uma das características extraídas do vídeo de entrada é repassada a um classificador. Esses classificadores utilizam o modelo gerado na etapa de Treinamento para apresentar sua sugestão (ou voto) de categoria a ser atribuída ao vídeo. Cada classificador deve optar por uma entre duas alternativas: a) positiva, indicando que o vídeo pertence à classe *stuffing* ou *spam*, de acordo com a situação em análise, ou; b) negativa, indicando que o vídeo representa um vídeo-resposta legítimo. O conjunto de votos gerado é analisado em uma etapa de combinação de evidências, que define em qual categoria o vídeo deve ser enquadrado.

4.2 Avaliação de Características Visuais: Atributos Globais × Atributos Locais

A motivação para a realização deste estudo é a existência, de resultados conflitantes, na literatura, sobre as vantagens relativas ao uso de descritores globais e locais em tarefas de classificação de informação em vídeos. Datta et al. [2008] observam que todos os métodos de representação de conteúdo visual têm vantagens e desvantagens.

Na tarefa específica de detecção de duplicatas, as primeiras técnicas eram baseadas em descrição global, como exemplificado por Chang et al. [1998], que utilizaram *wavelets* para representar o conteúdo visual aplicando na detecção de duplicatas de imagens na *web*, e Naphade e Yeo [2000] que utilizaram uma abordagem de intersecção de um conjunto de histogramas nos canais YUV, aplicada na detecção de duplicatas de vídeos. Posteriormente, as abordagens baseadas em descritores locais foram identificadas como mais vantajosas e se tornaram mais correntes, tal como realizado por Ke et al. [2004], que utilizaram o PCA-SIFT para descrever conteúdo visual aplicando na detecção de duplicatas e recuperação de vídeos, e Valle et al. [2008], que utilizaram o SIFT para descrever o conteúdo visual aplicado na recuperação de imagens.

Nas revisões elaboradas por Kherfi et. al. [2004] e Datta et al. [2008] são avaliadas diferentes técnicas globais e locais aplicadas em tarefas de recuperação de imagens por conteúdo e é indicado que os descritores locais alcançam melhores resultados na representação de detalhes ou quando se busca por um conteúdo específico dentro de uma imagem, tornando-os mais viáveis para o uso em tarefas de recuperação por conteúdo. Contudo, em problemas que necessitam da avaliação do conteúdo da imagem como um todo, técnicas globais apresentam melhor desempenho.

Para avaliar o desempenho no contexto da classificação de vídeos não colaborativos, foram selecionadas algumas técnicas representantes de cada abordagem. Adotou-se como critério o uso de técnicas consolidadas na área e eventuais propostas de adaptação que indicassem aplicabilidade em nosso domínio. Como representantes das técnicas globais, foram selecionados o histograma de matiz, histograma nos canais RGB e momentos de Zernike [Khotanzad e Hong, 1990]. As duas primeiras selecionadas por serem abordagens

tradicionais e a última, por apresentar bom desempenho em trabalhos recentes de reconhecimento de formas [Singh et al., 2011; Lajevardi e Hussain, 2009].

As abordagens locais selecionadas foram o SIFT [Lowe, 1999], PCA-SIFT [Ke e Sukthankar, 2004] e HUE-SIFT [Van de Weijer et al., 2006]. A primeira foi selecionada por ser a mais tradicional e as outras por serem variações que demonstraram bom desempenho na identificação de categorias semânticas, conforme observado por Wu et al. [2007], que aplicaram o PCA-SIFT na detecção de vídeos-duplicatas, e Van de Sande [2010], que utilizou o HUE-SIFT no reconhecimento de cenas em vídeos e verificou desempenho similar o SIFT.

4.3 Avaliação de Características Atributos Espaço-Temporais × Atributos Espaciais

Os atributos espaço-temporais foram desenvolvidos com o objetivo de representar com maior fidedignidade os conteúdos visuais presentes em vídeos, visto que os métodos puramente espaciais foram elaborados para o domínio de imagens. Como os demais atributos locais, os descritores espaço-temporais atuam com base em regiões de interesse identificadas no vídeo. Wang et al. [2009] apresentam comparações entre diversas abordagens para a detecção de pontos de interesse espaço-temporais. Ainda, avaliam o uso de diferentes técnicas para a descrição do conteúdo dessas regiões e observam que o detector e o descritor apresentado por Laptev [2005] são os que alcançam melhores resultados na maioria dos experimentos realizados. Laptev [2005] sugere o uso do *Histogram of Oriented Gradients*, Histograma de Orientação de Gradientes, concatenado com o *Histogram of Optic Flow*, Histograma de Fluxo Óptico, como descritor do conteúdo visual das regiões de interesse.

O uso de descritores espaço-temporais tem sido observado como mais vantajoso na tarefa de classificação de vídeos em categorias semânticas, tal como: Zhao e Pietikäinen [2007] os aplicam no reconhecimento de expressões faciais, enquanto Khoshabeh e Hollan [2009], Lopes et al. [2009c] e Shabani et al. [2009] os aplicam especificamente no reconhecimento de ações humanas. Contudo, no nosso trabalho, a aplicação se dá em um cenário distinto, com maior variabilidade visual extraclasses. Por isso, se justifica a

verificação, no cenário de aplicação deste trabalho, do desempenho de abordagens espaço-temporais em contraste com as puramente espaciais.

Os descritores locais puramente espaciais avaliados são: SIFT [Lowe, 1999], PCA-SIFT [Ke e Sukthankar, 2004] e HUE-SIFT [Van de Weijer et al., 2006]. Quanto à abordagem espaço-temporal foi escolhido o STIP [Laptev, 2005].

4.4 Avaliação de Características Multimodais: Combinação de Evidências Multimodais × Evidências Monomodais

A combinação de evidências multimodais se apresenta como uma alternativa promissora para o aumento na qualidade dos resultados na tarefa de classificação de dados multimídia na *web*. Brezeale e Cook [2008] elaboraram uma revisão de técnicas e atributos para a representação de vídeos e mostram que a principal dificuldade na realização dessa tarefa está na definição da melhor alternativa para a análise conjunta das informações oriundas de diferentes fontes. Existem abordagens que realizam a junção de todos os dados em uma característica única, *early fusion*, enquanto outras propõem o uso de classificadores independentes, e posterior fusão das evidências, na técnica chamada *late fusion*.

Em ambiente *web*, os vídeos não seguem regras rígidas quanto à composição ou apresentação de conteúdo. Foi observado que vídeos na *web* são caracterizados por uma maior diversidade de qualidade, assuntos, estilos e gêneros quando comparados a vídeos observados em outros ambientes [Yang et al., 2007a]. Essa característica faz com que a aplicação de técnicas de classificação em categorias semânticas seja mais complexa nesse ambiente. Por outro lado, geralmente, vídeos na *web* vêm acompanhados de informações que auxiliam na sua descrição e identificação no ambiente virtual. Jiang [2010] sugere que para a realização de uma classificação eficaz de vídeos na *web* em categorias semânticas, devem ser analisadas as meta-informações disponíveis sobre cada vídeo, além de suas características visuais.

Na literatura da área são observados outros trabalhos que utilizam atributos multimodais para a representação de imagens/vídeos como alternativa para ampliar a

qualidade dos resultados da classificação, como foi observado por: Lin e Hauptmann [2002] que utilizaram o classificador SVM aliado à combinação de evidências textuais, extraídas de legendas, e visuais para classificação de vídeos de jornalismo; Yan e Naphade [2005] que combinaram evidências textuais e audiovisuais, com base no classificador SVM, para classificação semântica supervisionada de vídeos; Liu et al. [2006] que utilizaram a combinação de evidências de audiovisuais para a classificação semântica de vídeos de jogos de basquete, e; Guillaumin et al. [2010] que combinaram informação visual e textual, obtida de palavras-chave, em um esquema de classificação semisupervisionada para categorização de imagens.

Neste experimento é sugerido o uso combinado de informações visuais e estatísticas em contraste ao uso de evidências puramente visuais. Trata-se de uma forma muito limitada de "multimodalidade", mas os resultados podem servir como motivador para trabalhos futuros de inclusão de informações mais ricas, provenientes de metadados textuais e áudio. As informações visuais são obtidas com o uso de descritores visuais de baixo-nível. As informações estatísticas são obtidas após a etapa de Pré-Processamento e demonstram uma medida da variabilidade do conteúdo visual presente no vídeo. Essa característica é composta pela informação de número de quadros que compõem o resumo do vídeo e o valor obtido da relação entre “quantitativo original de quadros” e “número de quadros-chave no resumo”, representando a noção de variabilidade de conteúdo visual presente no vídeo.

4.5 Avaliação de Características Visuais: Combinação de Evidências × Evidências Isoladas

Considerando que cada atributo descreve uma característica visual distinta, a adoção de diferentes evidências combinadas se apresenta como uma alternativa instigante em situações onde o conjunto de vídeos possui uma grande variabilidade visual. Na visão de Sharkey e Sharkey [1997] a combinação “ótima” de classificadores só pode ser obtida se esses apresentarem bom desempenho individual e ao mesmo tempo alta diversidade. Os aspectos cognitivos envolvidos no processo de tomada de decisão coletiva são explorados

por Coppin et al. [2007], auxiliando no entendimento da tarefa de combinação de evidências.

As estratégias de combinação de evidências adotadas neste trabalho são baseadas na abordagem *late fusion*. Ou seja, inicialmente são realizadas ações individuais, com classificadores independentes para cada uma das características. Em seguida, todos os resultados são utilizados como fonte de informação para se determinar a qual categoria um dado vídeo pertence.

A abordagem *late fusion* permite que se analise qual a influência de cada uma das características visuais na atribuição da classificação final. Com o uso de outra abordagem, a *early fusion*, esse tipo de análise se demonstra inviável, além de elevar a dimensionalidade do vetor de características. Contudo, segundo Polikar [2006] as técnicas *late fusion* são ainda um problema de pesquisa em aberto, em que a qualidade dos resultados está diretamente associada ao nível de controle que se tem do ambiente de aplicação. Devido às características do domínio de aplicação deste trabalho, tem-se um baixo nível de controle do ambiente. Isso aumenta o interesse em se verificar o desempenho da técnica *late fusion*.

Diferentes trabalhos na literatura observam a adoção da combinação de evidências, via *late fusion*, como sendo uma alternativa para aumentar a qualidade da classificação, exemplos recentes são: Le et al. [2007] que aplicam essa estratégia na definição do significado de palavras em uma base textual; Fagni et al. [2010] que propõem diferentes modelos adaptativos para a combinação de evidências, aplicando na classificação de imagens, e; Sannen et al. [2010] que propõem uma nova técnica para a combinação incremental, também aplicando em uma base de imagens.

Ainda, na literatura são encontradas diferentes abordagens de combinação de evidências *late fusion*, tais como: “mais votado” [Ruta e Gabrys, 2005], Redes Neurais [Cybenko, 1989], Modelo de Mistura Gaussianas [Paalanen et al. 1995], Modelos de Decisão [Kuncheva et al., 2001], *Boosting* [Freud e Schapire, 1996] e *Bagging* [Skurichina e Duin, 1998]. Neste trabalho é adotado o uso da técnica de “mais votado”, por verificar que esta é uma das mais tradicionais e promissoras, como observado nos estudos desenvolvidos por Ruta e Gabrys [2000, 2001, 2002 e 2005]. Nesses estudos é sugerido que com o uso da abordagem do “mais votado” é possível atingir resultados similares aos

de técnicas mais complexas. Ainda, por não se ter um conhecimento prévio quanto à complementaridade, ou não, dos descritores visuais utilizados, a adoção de uma técnica simples de combinação se apresenta como sendo mais indicado.

Neste experimento são avaliadas diferentes combinações de evidências visuais em contraste com o uso isolado dessas.

4.6 Avaliação de Classificadores: SVM × Naïve Bayes

Matsui e Tanabe [2006], Ulusoy e Bishop [2005], Ng e Jordan [2001], Joshi et al. [2005], Xue e Titterington [2008], Bíró et al. [2008] e Silovsky et al. [2009] elaboraram estudos comparativos entre técnicas de abordagem discriminativa e generativa, em diferentes contextos. Ulusoy e Bishop [2005] observaram que o modelo discriminativo é capaz de se concentrar nas características mais informativas, enquanto o modelo generativo demonstra maior precisão na classificação e melhor capacidade de localização de objetos na imagem. Ng e Jordan [2001] compararam o desempenho de classificadores discriminativos em contraste com os generativos, sendo observado que a opinião dominante na área indica que classificadores discriminativos são preferidos em detrimento aos generativos e que essa decisão não está associada à questões de custo computacional.

Classificadores generativos aprendem um modelo de probabilidade conjunta, $p(x, y)$, para uma entrada x e uma classe y , e realizam sua predição utilizando a regra de Bayes para calcular $p(y|x)$, e então selecionar a classe mais provável. Os classificadores discriminativos calculam a probabilidade *a posteriori* $p(y|x)$ diretamente.

Na literatura da área são encontradas diferentes propostas de métodos de classificação que se baseiam no modelo generativo, tais como: *GMM – Gaussian Mixture Models* (Modelos de Misturas de Gaussianas) [Permuter, 2003], *HMM – Hidden Markov Models* (Modelos Oculto de Markov) [Rabiner, 1989], *Naïve Bayes* (Bayes Ingênuo, em português) [Ng e Jordan, 2001], *LDA – Latent Dirichlet Allocation* (Alocação Latente de Dirichlet) [Blei et al., 2003], *pLSA – Probabilistic Latent Semantic Analysis* (Análise Probabilística de Semântica Latente) [Hofmann, 1999], entre outras.

Neste trabalho, os métodos SVM e *naïve bayes* foram selecionados como sendo representantes das abordagens discriminativa e generativa, respectivamente, por serem técnicas mais tradicionais na literatura.

4.7 Conclusões

Esse capítulo apresentou a solução modelo que possibilita o uso isolado, ou combinado, de informação visual advindas de diferentes descritores. Com essa solução modelo é possível avaliar a validade, ou não, de várias das hipóteses levantadas neste trabalho. No próximo capítulo, são apresentados os experimentos realizados e os resultados obtidos com a execução do projeto experimental que utiliza a solução modelo.

Capítulo 5

Abordagens tradicionais —

Experimentos e Resultados

No presente capítulo serão apresentadas a descrição das bases de dados utilizadas e os experimentos realizados na avaliação de diferentes hipóteses, utilizando a solução modelo descrita no Capítulo 4. Uma dificuldade inerente do escopo escolhido para aplicação deste trabalho é a falta de uma referência para comparação dos resultados obtidos e posicionamento destes em relação aos demais trabalhos similares desenvolvidos na área. Devido a isso, foram realizados experimentos com técnicas comuns nas áreas correlatas para servirem de referência. Os experimentos foram divididos de acordo com o tipo de ação não colaborativa analisada, *stuffing* ou *spam*, e a hipótese em avaliação.

5.1 Aquisição da Base de Dados

Para uma melhor descrição da base de dados envolvida neste grupo de experimentos, a Subseção 5.1.1, apresenta o conjunto de dados, chamado de **selvagem**, que consiste em uma base experimental obtida através de uma amostra representativa de casos encontrados na rede social.

Os dados utilizados para realizar a classificação automática foram extraídos de características visuais e, em um experimento específico, de análise estatística da estrutura dos próprios vídeos. Não foram usadas outras informações que porventura possam estar associadas ao vídeo, como: descrição textual, nome de usuário, tempo de duração, categoria associada, comentários de usuários, entre outras. Assim, não se analisou qualquer outra informação da rede social se não os próprios vídeos, a informação acerca

de a qual *thread* o vídeo pertence e se este é o vídeo gerador da *thread* (vídeo original) ou um vídeo-resposta.

5.1.1 Base de Dados “Selvagem”

Os vídeos foram coletados no YouTube no período compreendido entre maio de 2009 e janeiro de 2010. Essa rede social foi selecionada por ser a de maior popularidade entre as comunidades virtuais para compartilhamento de vídeo digital¹² e por ser pioneira no conceito de *threads* de vídeo-reposta. A construção da base de dados não seguiu critérios de rede de relacionamentos (entre usuários, etc.) para a coleta de vídeos. Ao invés disso, procurou-se coletar *threads* em que os problemas de *spam* e *stuffing* estivessem mais presentes.

O YouTube disponibilizava, naquele momento, uma relação com os vídeos que apresentam maior volume de vídeos-resposta, “lista de mais respondidos”. Foi utilizada essa lista como ponto de partida para a identificação das *threads* e coleta dos vídeos-resposta para comporem a base de dados experimental. Todos os vídeos-resposta das *threads* selecionadas foram coletados, utilizando uma aplicação¹³, em *Python*, para auxiliar na automatização da coleta dos vídeos.

Foram coletados 11.360 vídeos, incluindo originais e resposta, selecionados de um conjunto de 83 *threads*, identificadas aleatoriamente na lista de mais respondidos. Esses vídeos foram submetidos a uma etapa de anotação manual, realizada pelo autor do presente trabalho. Deve ser ressaltado que devido à etapa de anotação ter sido realizada por apenas uma pessoa, há a possibilidade de presença de viés nesse processo. Contudo, há circunstâncias amenizantes, como o fato dos vídeos terem sido anotados *antes* da concepção das soluções apresentadas. É importante ter em mente a dificuldade e o custo de execução dessa etapa. Ainda, é importante ressaltar que, mesmo tendo sido coletado um bom volume de vídeos, devido o grande número e diversidade de vídeos disponíveis na rede social, a dinamicidade do aumento desse volume e a dificuldade para anotação manual dos vídeos, o conjunto de vídeos coletados e utilizados neste trabalho não representa exaustivamente a real diversidade e dinâmica de uma rede social.

¹² Conforme observado em consulta ao site www.alexacom.com em fevereiro de 2012.

¹³ *YouTube-dl*, disponível em <http://bitbucket.org/rg3/YouTube-dl/wiki/Home>.

Definir automaticamente que um vídeo tem a intenção de manipular a popularidade de uma *thread* de vídeos-resposta não é uma tarefa trivial. Algumas situações observadas podem causar incerteza durante a categorização: “qual o nível de similaridade entre os vídeos para caracterizar a igualdade no tipo *stuffing*?”, “quantos vídeos iguais são necessários existir em uma *thread* para que seja caracterizada a ocorrência de vídeos do tipo *stuffing*?”, “se em um vídeo original é apresentado um clipe musical de um dado artista, os vídeos-resposta contendo clipes musicais de outros artistas são similares a este?”, “se em um vídeo original é apresentada uma entrevista acerca de um tema polêmico, um vídeo-resposta que contenha uma cena de pessoas manifestando acerca do mesmo tema é considerado ‘legítimo’?”.

Essas são apenas algumas questões que representam a dificuldade para a anotação manual dos vídeos-resposta. Dada à complexidade que o problema impõe, foram definidos alguns critérios rígidos para a atribuição de uma categoria a um dado vídeo. Definiu-se que: a) dois ou mais vídeos com algum conteúdo similar, que não apresentem relação com a informação contextual observada no vídeo original, são caracterizados como *stuffing*; b) um vídeo que possua algum conteúdo visual similar com o vídeo original ou com qualquer outro vídeo-resposta previamente classificado como “legítimo”, é classificado como “legítimo”; c) os demais vídeos são classificados como *spam*. E em caso de dúvida por parte do anotador, adotou-se uma postura conservadora, considerando os vídeos como “legítimo”, mesmo caminho seguido por Benevenuto (2009).

A Tabela 5.1 apresenta como estão distribuídos os vídeos nas classes definidas no contexto deste trabalho: “original”, *spam*, *stuffing* e “legítimo”.

Tabela 5.1. Quantidade de vídeos coletados por classe.

Categorias	Número de Vídeos
Original	83
Legítimo	3.420
<i>Spam</i>	4.678
<i>Stuffing</i>	3.179

Para reduzir a possibilidade de viés na classificação, garantindo o equilíbrio entre os elementos das diferentes classes, e reduzir o custo computacional, os experimentos deste trabalho foram executados considerando apenas um subconjunto dos vídeos coletados,

aleatoriamente escolhidos. A Tabela 5.2 apresenta a quantidade de vídeos-resposta em cada categoria e o número de quadros extraídos desses vídeos. O número de quadros apresentados considera apenas os elementos resultantes da etapa de pré-processamento que converteu os vídeos em imagens estáticas, amostrando os quadros de acordo com a taxa de quadros por segundo de cada vídeo.

Tabela 5.2. Distribuição de vídeos e quantitativo de quadros por classe, utilizados nos experimentos deste trabalho.

Categorias	Experimento	Número de Vídeos	Número de Quadros
Legítimo	Legítimo x <i>Spam</i>	1.000	169.956
<i>Spam</i>		1.000	158.622
Legítimo	Legítimo x <i>Stuffing</i>	500	84.486
<i>Stuffing</i>		500	8.782

Analisando os valores apresentados na Tabela 5.2, pode ser compreendida a complexidade da manipulação das informações em todos os experimentos. E, dada à variabilidade do conteúdo visual apresentado nos vídeos das classes *spam* e “legítimo”, o experimento que se propõe a distinguir os vídeos dessas classes recebeu um conjunto maior de elementos.

5.2 Visão Geral dos Experimentos

Foi observado em testes preliminares e demais trabalhos de classificação encontrados na literatura que, considerando a diversidade de comportamento dos elementos das categorias semânticas contempladas neste trabalho (legítimo, *spam* e *stuffing*), uma subdivisão dos experimentos em duas etapas facilitaria a análise dos resultados obtidos. Cada etapa avaliou o comportamento de uma categoria não colaborativa, *spam* e *stuffing*, em relação aos vídeos legítimos. Com isso, os experimentos analisaram de modo separado as relações entre *Stuffing* × *Legítimo* e *Spam* × *Legítimo*.

Foram analisados atributos visuais que atuam de modo global e local: a) *global*: histograma nos canais RGB e histograma de matiz, que estão relacionados à informação de cor presente no quadro, e; momentos de Zernike de 10^a ordem, que estão relacionados às formas de objetos contidos no quadro em análise; b) *local*: *SIFT*, *PCA-SIFT* e *HUE-SIFT*,

que atuam em quadros do vídeo de modo isolado, descrevendo informações puramente espaciais, e; *STIP*, que considera todos os quadros presentes no vídeo bem como a dependência entre eles, descrevendo informação espaço-temporal. Além dos atributos visuais, é adicionada uma característica composta por duas informações obtidas após a etapa de pré-processamento dos vídeos, nomeada de “estatísticas de quadro”, conforme visto na Seção 4.4.

Além dos atributos individuais, foram analisadas as seguintes combinações: globais, que reúne as evidências individuais de cada um dos atributos globais extraídos no contexto deste trabalho; locais, que é constituído com base na combinação de evidências de todos os atributos locais; todos os visuais, que reúne as combinações Globais e Locais em uma só combinação; as evidências do atributo *STIP* com as estatísticas de quadro; as evidências do atributo *SIFT* com as estatísticas de quadro, e; as evidências do atributo histograma de matiz com as estatísticas de quadro.

Nas próximas seções são apresentados os resultados obtidos nos experimentos realizados. O classificador SVM com *kernel* linear foi adotado em todos os experimentos, com exceção de onde é especificamente descrito o uso de outro classificador. Os dados de entrada são normalizados antes da execução do classificador, utilizando uma escala $[-1, +1]$. Como estratégia de validação dos resultados é adotada a validação cruzada (*cross-validation*, em inglês) com 5 dobras (*folds*, em inglês). Os instrumentos para avaliação da qualidade dos resultados da classificação são a Taxa de Verdadeiro Positivo (TPR), Taxa de Falso Positivo (FPR) e Coeficiente de Correlação de Matthews (MCC). A ANOVA (do inglês *Analysis of Variance*, Análise de Variâncias) e o teste-t par-a-par são utilizados para verificar a significância estatística dos experimentos.

O MCC é uma medida agregada que mensura a qualidade do resultado da detecção, considerando a correlação entre as classes esperadas e as preditas pelo classificador. O valor é apresentado em uma escala de -1 a +1, sendo que: +1 representa uma predição perfeita; o valor 0 corresponde a uma classificação aleatória, e; o valor -1 indica uma predição completamente equivocada.

Por ter o seu resultado influenciado pelos valores de TPR e FPR conjuntamente, o MCC foi selecionado para ser avaliado na ANOVA e no teste-t. Na Tabela 5.3 são apresentadas as formulações para obtenção dos valores de TPR, FPR e MCC.

O limite adotado para rejeição da hipótese-nula na ANOVA foi $\alpha = 0,05$ (nível de confiança de 95%). Em todos os casos a hipótese-nula é de que não existe diferença significativa entre os valores de MCC das populações avaliadas. As avaliações da ANOVA e teste-t são complementares: enquanto a ANOVA avalia as diferenças de médias entre as populações e mede a significância estatística do conjunto de experimentos como um todo, o teste-t avalia a diferença da média entre pares de experimentos identificando a significância, ou não, no comparativo individual entre cada uma das configurações avaliadas.

Tabela 5.3. Fórmulas para o cálculo de Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC).

$$TPR = \frac{TP}{(TP + FN)}$$

$$FPR = \frac{FP}{(FP + TN)}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

Legenda:
TP: Verdadeiros Positivos (*True Positives*, em inglês).
TN: Verdadeiros Negativos (*True Negatives*, em inglês).
FN: Falsos Negativos (*False Negatives*, em inglês).
FP: Falsos Positivos (*False Positives*, em inglês).

5.2.1 Experimentos *Stuffing* × *Legítimo*

Os experimentos realizados nesse grupo utilizam o conjunto de 1.000 vídeos-resposta selecionado aleatoriamente na base **selvagem**. As classes são representadas por 500 vídeos da classe “legítimo” e 500 vídeos da classe *stuffing*.

A categoria *stuffing* está associada à classificação positiva e a categoria legítimo à classificação negativa. Ou seja, durante a avaliação dos resultados da classificação a identificação correta de um vídeo do tipo *stuffing* é considerada como verdadeiro positivo, enquanto que a identificação correta de um elemento da classe legítimo é considerada como verdadeiro negativo.

5.2.2 Experimentos *Spam* × Legítimo

Nos experimentos deste tipo é utilizado o conjunto de 2.000 vídeos-resposta selecionado aleatoriamente na base **selvagem**. As classes são representadas por 1.000 vídeos da classe legítimo e 1.000 vídeos da classe *spam*. O volume de vídeos utilizados é maior que o experimento anterior, devido observar-se uma grande dificuldade na separação entre os elementos de cada classe durante a etapa de anotação manual. O aumento no volume de vídeos tem a intenção de proporcionar o aprendizado de mais elementos de cada classe, já que esta tarefa é mais difícil.

A categoria *spam* está associada à classificação positiva e a categoria legítimo à classificação negativa. Durante a avaliação dos resultados da classificação a identificação correta de um vídeo *spam* é considerada como verdadeiro positivo, enquanto que a identificação correta de um elemento da classe legítimo é considerada como verdadeiro negativo.

5.3 Experimentos e Resultados

Utilizando a solução modelo, apresentada no Capítulo 4, é procedida com a verificação da validade de algumas das hipóteses de pesquisa levantadas neste trabalho, conforme segue:

- **Hipóteses 1.1 e 1.2:** na Seção 5.4, são avaliados os diferentes descritores visuais, possibilitando uma análise comparativa do desempenho das diferentes abordagens: global x local (puramente espacial), e; espacial x espaço-temporal;
- **Hipótese 3:** na Seção 5.5, é analisado o desempenho do uso combinado das evidências visuais, utilizando uma estratégia *late fusion* baseada em votação simples;
- **Hipótese 2:** na Seção 5.6, é analisado o desempenho de diferentes abordagens de classificação supervisionada, avaliando o uso dos classificadores SVM e Naïve Bayes;

- **Hipótese 1.3:** na Seção 5.7, é analisado o desempenho de abordagem de combinação de evidências multimodal, utilizando os descritores visuais e “estatísticas de quadro”.

5.4 Avaliando descritores visuais em atuação isolada

Para avaliar o desempenho dos diferentes descritores visuais, em atuação isolada frente ao problema da detecção de vídeos não colaborativos, foram realizados experimentos com descritores: a) **globais**: histograma de matiz, histograma nos canais RGB e momentos de Zernike; b) **locais, puramente espaciais**: SIFT, PCA-SIFT e HUE-SIFT, e; c) **local, espaço-temporal**: STIP, com descritor HOG-HOF [Laptev, 2005]. Todos os descritores utilizam a distância euclidiana como métrica de similaridade.

Na Tabela 5.4 são apresentadas as médias dos valores de Taxas de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC) obtidas nos experimentos de classificação utilizando cada um dos descritores visuais nos conjuntos de dados *Stuffing* × Legítimo (a) e *Spam* × Legítimo (b).

A ANOVA obteve um *p-value* <0,0001 para os experimentos de ambos os conjuntos de dados, com isso é rejeitada a hipótese-nula e as diferenças das médias dos valores de MCC são consideradas como significativas tanto para os experimentos *Stuffing* × Legítimo quanto *Spam* × Legítimo, com um nível de confiança de 95%. Na Tabela 5.5 são apresentados os resultados para os testes par-a-par entre os experimentos.

Tabela 5.4. Valores médios para Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC). Os melhores resultados individuais aparecem destacados em negrito.

Stuffing × Legítimo					Spam × Legítimo			
		TPR	FPR	MCC		TPR	FPR	MCC
Hist. Matiz	1	0,604	0,138	0,472	1	0,954	0,913	0,083
Hist. RGB	2	0,632	0,220	0,422	2	0,907	0,843	0,097
Zernike	3	0,408	0,070	0,395	3	0,948	0,924	0,051
SIFT	4	0,608	0,110	0,511	4	0,490	0,339	0,152
PCA-SIFT	5	0,602	0,178	0,432	5	0,750	0,600	0,163
Hue-SIFT	6	0,300	0,140	0,388	6	0,797	0,507	0,298
STIP	7	0,980	0,126	0,816	7	0,679	0,583	0,099

Tabela 5.5. Matriz com resultados do teste-t para os valores de Coeficiente de Correlação de Matthews (MCC) com significância de 95%.

		Stuffing × Legítimo							Spam × Legítimo							
		1	2	3	4	5	6	7		1	2	3	4	5	6	7
Hist. Matiz	1		x	+	x	x	○	.	1		x	x	+	○	.	x
Hist. RGB	2	x		x	○	x	x	.	2	x		x	x	+	.	x
Zernike	3	+	x		●	x	x	.	3	x	x		●	●	.	x
SIFT	4	x	○	●		+	●	.	4	+	x	●		x	.	x
PCA-SIFT	5	x	x	x	+		x	.	5	○	+	●	x		.	+
Hue-SIFT	6	○	x	x	●	x		.	6	.	.	.	.	.		.
STIP	7	.	.	.	.	.	.		7	x	x	x	x	+	.	

Legenda:

Rejeição de Hipótese Nula

< 0,001 .

< 0,01 ●

< 0,05 ○

Não Rejeição de Hipótese Nula

< 0,1 +

>= 0,1 x

Na avaliação de desempenho dos descritores globais em relação aos locais puramente espaciais, no conjunto *Stuffing* × Legítimo, é observado que os atributos com melhores resultados são histograma de matiz (global), SIFT (local) e PCA-SIFT (local). Os resultados obtidos são muito próximos em todos os três critérios de análise: a) histograma de matiz: 0,604 (TPR); 0,138 (FPR), e; 0,472 (MCC); b) SIFT: 0,608 (TPR); 0,110 (FPR), e; 0,511 (MCC); c) PCA-SIFT: 0,602 (TPR); 0,178 (FPR), e; 0,432 (MCC). E, observando os resultados do teste par-a-par, a pequena vantagem numérica no MCC do SIFT não passa no teste de significância estatística, não sendo possível, portanto, afirmar que a diferença entre os descritores seja estatisticamente significativa com 95% de confiança.

Analisando o desempenho do descritor espaço-temporal (TPR: 0,980; FPR: 0,126; MCC: 0,816) é observada uma enorme vantagem para os demais com relação aos valores de TPR e MCC. E, na avaliação do teste par-a-par é demonstrada a rejeição da hipótese-nula para todas as comparações, permitindo tomar como verdadeira a superioridade desse descritor, com 95% de significância.

As características dos vídeos que compõem a categoria *stuffing* (curtíssima duração; pouca variabilidade de conteúdo entre os quadros, e; presença de vários vídeos com conteúdo visual muito similar, variando apenas no texto que apresenta ou cor de fundo do quadro) contribuem para o bom desempenho de técnicas globais. Ainda, as características

do *stuffing* e o fato dos vídeos legítimos geralmente apresentarem maior duração e variabilidade de conteúdo visual, dão margem para que abordagens que consideram a informação temporal, tal como o STIP, tenham bom desempenho na detecção de vídeos do tipo *stuffing*.

Um fato interessante é o baixo desempenho do atributo HUE-SIFT. Por se tratar de uma variação do SIFT que adiciona informações de cor do canal de matiz em sua descrição, esperava-se que obtivesse resultado similar aos alcançados por SIFT e histograma de matiz.

As características visuais analisadas na classificação entre as categorias *spam* e legítimo obtiveram, no geral, resultados insatisfatórios, próximos da linha do acaso. O baixo desempenho de todos os descritores nesse experimento pode ser atribuído a grande variabilidade de conteúdo visual intraclasse observada nos elementos de ambas as categorias. Outro fator capaz de influenciar o resultado é o “pequeno” volume de dados para a constituição da base de treinamento. Contudo, deve-se observar que não é possível saber *a priori* qual o tamanho ideal para que o conjunto de treinamento obtenha um aumento significativo na qualidade dos resultados. O grande volume de vídeos disponíveis na comunidade virtual e a liberdade e criatividade que um usuário possui para geração de vídeos são complicadores ainda distantes de serem minimizados.

5.5 Avaliando descritores visuais em atuação conjunta

Foram elaborados três conjuntos de combinações de descritores, utilizando a estratégia *late fusion* de votação simples: a) globais: composto por histograma de matiz, histograma nos canais RGB e momentos de Zernike; b) locais: composto por SIFT, PCA-SIFT, HUE-SIFT e STIP, e; c) todos os visuais: composto pelas combinações globais e locais. Para avaliar o desempenho das diferentes combinações de descritores foram selecionados três dentre os melhores resultados individuais para servirem de referência: histograma de matiz (global), SIFT (local, puramente espacial) e STIP (local, espaço-temporal).

Tabela 5.6. Valores médios para Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC). Os melhores resultados individuais aparecem destacados em negrito.

Stuffing × Legítimo					Spam × Legítimo				
		TPR	FPR	MCC		TPR	FPR	MCC	
Hist. Matiz	1	0,604	0,138	0,472	1	0,954	0,913	0,083	
SIFT	2	0,608	0,110	0,511	2	0,490	0,339	0,152	
STIP	3	0,980	0,126	0,816	3	0,679	0,583	0,099	
Globais	4	0,626	0,086	0,547	4	0,982	0,938	0,114	
Locais	5	0,606	0,016	0,619	5	0,744	0,407	0,331	
Todos Visuais	6	0,678	0,028	0,658	6	0,968	0,864	0,186	

Os *p-values* obtidos com a ANOVA foram $<0,0001$ para os experimentos *Stuffing* × Legítimo quanto *Spam* × Legítimo, com isso são rejeitadas as hipóteses-nulas e as diferenças das médias dos valores de MCC são consideradas como significativas, com um nível de confiança de 95%. Na Tabela 5.6 são apresentados os valores médios para Taxas de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC), obtidos nos experimentos utilizando os descritores selecionados e as combinações elaboradas. Na Tabela 5.7, são apresentados os resultados para os testes par-a-par, com um nível de confiança de 95%.

Nos resultados da classificação entre elementos das classes *stuffing* e legítimo pode ser observado que nenhuma das combinações experimentadas obteve resultado superior ao uso do atributo STIP, considerando os fatores de TPR e MCC. Nas análises de significância estatística entre cada uma das combinações e atributo STIP são rejeitadas as hipóteses-nulas de que não existe diferença entre os resultados, o que valida a vantagem na qualidade dos resultados do STIP em relação aos demais. Contudo, vale ressaltar que as combinações demonstram melhores resultados na identificação positiva de vídeos legítimos, indicado pela FPR: globais (0,086), locais (0,016) e todos visuais (0,028).

Com os resultados desse experimento, é possível a utilização de uma combinação, tal como a locais, para construção de sistemas que primam pela identificação positiva dos elementos legítimos, entendendo que é menos prejudicial apresentar um número maior de vídeos *stuffing* e reduzir o risco de deixar de fora um vídeo legítimo.

Assim como observado no desempenho do uso dos descritores isolados, a utilização combinada de descritores não apresentou resultados satisfatórios para os experimentos

no conjunto *Spam* × Legítimo, ficando muito longe do esperado para o uso em ambiente real.

Tabela 5.7. Matriz com resultados do teste-t para os valores de Coeficiente de Correlação de Matthews (MCC) com significância de 95%.

		Stuffing × Legítimo						Spam × Legítimo							
		1	2	3	4	5	6			1	2	3	4	5	6
Hist. Matiz	1		×	•	+	•	•		1		○	×	×	•	●
SIFT	2	×		•	×	●	•		2	○		+	×	•	×
STIP	3	•	•		•	•	•		3	×	+		×	•	●
Globais	4	+	×	•		+	●		4	×	×	×		•	○
Locais	5	•	●	•	+		×		5	•	•	•	•		•
Todos Visuais	6	•	•	•	●	×			6	●	×	●	○	•	

Legenda:

Rejeição de Hipótese Nula

< 0,001 .

< 0,01 ●

< 0,05 ○

Não Rejeição de Hipótese Nula

< 0,1 +

>= 0,1 ×

5.6 Avaliando classificadores SVM e Naïve Bayes

Para avaliar o desempenho da adoção do classificador *naïve bayes* foram selecionados três dentre os melhores resultados do uso de descritores individuais e classificador SVM com *kernel* linear para servirem de referência: histograma de matiz (global), SIFT (local, puramente espacial) e STIP (local, espaço-temporal). Ressalte-se que técnicas generativas mais sofisticadas que o *naïve bayes* poderiam ter obtido melhores resultados, embora o emprego satisfatório dessas técnicas exija grandes conjuntos de treinamento.

Na Tabela 5.8 são apresentadas as médias de Taxas de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC), obtidos nos experimentos utilizando os mesmos descritores visuais em ambos os classificadores. E, na Tabela 5.9, são apresentados os resultados para os testes par-a-par, com um nível de confiança de 95%. A ANOVA para esses experimentos apresentou *p-value* <0,0001 (*Stuffing* × Legítimo) e =0,0005 (*Spam* × Legítimo), rejeitando a hipótese-nula em ambas as situações e demonstrando a significância estatística da diferença entre as médias dos valores de MCC nos conjuntos.

Os resultados dos experimentos no conjunto *Stuffing* × *Legítimo* demonstram que em todos os casos o uso SVM demonstrou MCC superior aos demais, indicando maior qualidade na predição desse classificador. Mas, na análise par-a-par entre os experimentos utilizando SVM e *naïve bayes* considerando o mesmo descritor visual, apenas para o descritor Histograma de Matiz a hipótese-nula foi rejeitada. Com isso, não é possível afirmar que a superioridade obtida pelo SVM utilizando os outros descritores é estatisticamente significativa, com confiança de 95%.

Para o conjunto *Spam* × *Legítimo* a substituição do classificador não foi suficiente para obtenção de desempenho satisfatório.

Tabela 5.8. Valores médios para Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC). Os melhores resultados individuais aparecem destacados em negrito.

Stuffing × Legítimo					Spam × Legítimo			
		TPR	FPR	MCC		TPR	FPR	MCC
Hist. Matiz	1	0,604	0,138	0,472	1	0,954	0,913	0,083
SIFT	2	0,608	0,110	0,511	2	0,490	0,339	0,152
STIP	3	0,980	0,126	0,816	3	0,679	0,583	0,099
*NB – H. Matiz	4	0,820	0,692	0,153	4	0,119	0,113	0,010
*NB – SIFT	5	0,490	0,040	0,500	5	0,582	0,427	0,157
*NB – STIP	6	0,932	0,150	0,747	6	0,841	0,734	0,114

* Os experimentos precedidos por “NB” correspondem aqueles em que foi utilizado o classificador Naïve Bayes em substituição ao SVM.

Tabela 5.9. Matriz com resultados do teste-t para os valores de Coeficiente de Correlação de Matthews (MCC) com significância de 95%.

Stuffing × Legítimo								Spam × Legítimo							
		1	2	3	4	5	6			1	2	3	4	5	6
Hist. Matiz	1		×	•	•	×	•	1		○	×	○	○	○	×
SIFT	2	×		•	•	×	•	2	○		+	•	×	×	×
STIP	3	•	•		•	•	×	3	×	+		●	+	×	×
*NB – H. Matiz	4	•	•	•		•	•	4	○	•	●		•	●	●
*NB – SIFT	5	×	×	•	•		•	5	○	×	+	•			×
*NB – STIP	6	•	•	×	•	•		6	×	×	×	●	×		

Legenda:

Rejeição de Hipótese Nula

< 0,001 .

< 0,01 ●

< 0,05 ○

Não Rejeição de Hipótese Nula

< 0,1 +

>= 0,1 ×

* Os experimentos precedidos por “NB” correspondem aqueles em que foi utilizado o classificador Naïve Bayes em substituição ao SVM.

5.7 Avaliando a combinação de evidências multimodais

Para avaliar o desempenho da combinação de evidências multimodais foram selecionados três dentre os melhores resultados do uso de descritores individuais para servirem de referência: histograma de matiz (global), SIFT (local, puramente espacial) e STIP (local, espaço-temporal). Para cada descritor foi gerada uma combinação com a característica de “estatísticas de quadro”, que representa uma medida da variabilidade visual apresentada nos vídeos: a) Histograma de Matiz + estatísticas de quadro; b) SIFT + estatísticas de quadro, e; c) STIP + estatísticas de quadro.

Esse experimento foi idealizado para verificar se a junção de evidências visuais com informações oriundas de outras fontes pode, ou não, melhorar a qualidade dos resultados. Contudo, trata-se de uma forma muito limitada de multimodalidade, mas os resultados podem servir como motivador para trabalhos futuros de inclusão de informações mais ricas, provenientes de metadados textuais e áudio.

Na Tabela 5.10 são apresentadas as Taxas de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC), obtidos nos experimentos de classificação nos diferentes conjuntos, *Stuffing* × Legítimo (a) e *Spam* × Legítimo (b). E, na Tabela 5.11, são apresentados os resultados para os testes par-a-par, com um nível de confiança de 95%.

No conjunto *Stuffing* × Legítimo, as combinações utilizando os atributos histograma de matiz e SIFT apresentaram melhoria em todos os fatores analisados se comparado ao uso isolado destes descritores. Contudo, analisando o teste par-a-par de cada descritor visual com sua combinação multimodal correspondente, é observado que em nenhuma situação o teste de significância estatística rejeitou a hipótese-nula, o que indica que não há uma diferença significativa na qualidade da predição do classificador (indicada pelo valor de MCC), com confiança de 95%.

Já para o conjunto *Spam* × Legítimo a classificação utilizando a combinação multimodal não possibilitou a obtenção de desempenho satisfatório.

Tabela 5.10. Valores médios para Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Coeficiente de Correlação de Matthews (MCC). Os melhores resultados individuais aparecem destacados em negrito.

Stuffing × Legítimo					Spam × Legítimo				
		TPR	FPR	MCC		TPR	FPR	MCC	
Hist. Matiz	1	0,604	0,138	0,472	1	0,954	0,913	0,083	
SIFT	2	0,608	0,110	0,511	2	0,490	0,339	0,152	
STIP	3	0,980	0,126	0,816	3	0,679	0,583	0,099	
Estat. Quadro	4	0,960	0,214	0,727	4	0,456	0,353	0,106	
*Hist. Matiz +	5	0,674	0,094	0,580	5	0,875	0,762	0,147	
*SIFT +	6	0,706	0,094	0,609	6	0,716	0,424	0,289	
*STIP +	7	0,960	0,100	0,821	7	0,251	0,121	0,164	

* Os experimentos com o símbolo “+” ao final do nome correspondem aqueles em que foi utilizada a combinação de evidências da informação visual e “estatísticas de quadro”.

Tabela 5.11. Matriz com resultados do teste-t para os valores de Coeficiente de Correlação de Matthews (MCC) com significância de 95%.

Stuffing × Legítimo									Spam × Legítimo								
		1	2	3	4	5	6	7		1	2	3	4	5	6	7	
Hist. Matiz	1		×	•	•	○	●	•	1		○	×	×	○	•	○	
SIFT	2	×		•	•	+	○	•	2	○		+	×	×	•	×	
STIP	3	•	•		○	•	•	×	3	×	+		×	×	•	○	
Estat. Quadro	4	•	•	○		•	●	○	4	×	×	×		×	•	+	
*Hist. Matiz +	5	○	+	•	•		×	•	5	○	×	×	×		•	×	
*SIFT +	6	●	○	•	●	×		•	6	•	•	•	•	•		•	
*STIP +	7	•	•	×	○	•	•		7	○	×	○	+	×	•		

Legenda:

Rejeição de Hipótese Nula

< 0,001 •

< 0,01 ●

< 0,05 ○

Não Rejeição de Hipótese Nula

< 0,1 +

>= 0,1 ×

* Os experimentos com o símbolo “+” ao final do nome correspondem aqueles em que foi utilizada a combinação de evidências da informação visual e “estatísticas de quadro”.

5.8 Considerações

Com os experimentos apresentados nesse capítulo, algumas das hipóteses levantadas neste trabalho foram avaliadas (Hipóteses 1 a 3). Com a execução do experimento descrito na Seção 5.4, relativo à Hipótese 1.1, foi observado que o uso de técnicas de extração de características com abordagem global, em geral, apresentam resultados inferiores ao uso

de técnicas locais. Deve ser ressaltado que, no problema do *stuffing*, o histograma de matiz demonstrou resultado similar ao melhor descritor local puramente espacial.

Já na Seção 5.5, avaliando a validade da Hipótese 1.2, outra constatação, também com relação ao problema do *stuffing*, é a de que o uso de técnicas espaço-temporais amplia significativamente a qualidade dos resultados em comparação a adoção de estratégias puramente espaciais.

Na Seção 5.6, que descreve os experimentos referentes à Hipótese 5.6, uma observação importante é a de que, em geral, o uso do classificador bayesiano não apresentou resultados mais expressivos que o SVM, o que reforça a adoção do classificador SVM como padrão para execução dos experimentos.

Com os experimentos relativos às Hipóteses 1.3 e 3, descritos na Seção 5.7, pôde-se observar que as informações de “estatísticas de quadro” desempenharam papel importante na descrição dos elementos no conjunto *Stuffing* × Legítimo. Tal fato pode ser explicado pela distinção entre as características gerais dos vídeos que compõem cada uma das categorias analisadas, como a duração média e variabilidade de conteúdo visual dos vídeos. Essa constatação pode servir de motivador para o desenvolvimento de experimentos que combinem a descrição visual com evidências oriundas de outras fontes. Ainda, foi observado que a combinação de múltiplas evidências visuais pode ser utilizada em aplicações que estejam dispostas a aceitar uma menor taxa de detecção de vídeos positivos em troca de uma menor taxa de falsos alarmes.

Os experimentos e os resultados apresentados nesse capítulo demonstram que a identificação de elementos entre as categorias *Stuffing* e Legítimo alcançou resultados vistos como suficientes para justificar a utilização da solução modelo em ambientes reais. E, que a classificação entre as categorias *spam* e legítimo não obteve resultado satisfatório, sendo necessário avaliar outras abordagens para essa tarefa, como apresentado no Capítulo 6.

Como apenas o uso das características visuais de baixo nível semântico, mesmo aliado à técnicas de combinação de evidências, não foi suficiente para obtenção de uma abordagem eficaz para a tarefa de identificação de vídeos não colaborativos do tipo *spam*, no próximo capítulo é apresentada uma nova metodologia idealizada para utilizar outras abordagens, incluindo o contexto do vídeo e a projeção do vetor de características no

espaço de semântica latente. Os experimentos realizados com esse novo modelo utilizam apenas as características visuais SIFT e STIP. Essa escolha se deve ao fato de: a) o histograma de matiz, melhor descritor global, ter apresentado resultado enviesado no conjunto *Spam* $\times$ Legítimo; b) o SIFT ter demonstrado resultados iguais ou superiores aos demais descritores locais, de abordagem espacial, e; c) o STIP ter demonstrado resultado superior a todas as demais estratégias na maioria dos experimentos apresentados nesse capítulo.

Capítulo 6

Abordagem sensível ao contexto

O projeto experimental descrito no Capítulo 4 e avaliado no Capítulo 5 demonstrou efetividade na detecção do *stuffing*. Contudo, o mesmo ferramental não foi eficaz na identificação de *spam*. Isso demonstra que os problemas possuem graus de complexidade distintos e suscita a necessidade de uma estratégia específica para o problema de identificação do *spam*.

Os resultados obtidos no Capítulo 5 demonstram que não é viável estabelecer um padrão para o conteúdo visual do *spam* utilizando estratégias tradicionais. Um fator, de grande relevância, é que a condição de *spam* é dependente do tema discutido na *thread* na qual está inserido o vídeo, gerando uma dependência contextual nessa classificação, que não é tratada nas abordagens tradicionais. Outro fator é a grande variabilidade visual intraclasse, sendo praticamente impossível gerar um conjunto de treinamento que represente todas as possibilidades, seja para *spam* seja para legítimo.

Para enfrentar esses problemas propõe-se adicionar a informação de contexto do vídeo junto à representação de seu conteúdo visual. Utilizou-se um vídeo-referência, específico para cada *thread*, responsável por representar a informação do contexto na qual o vídeo está inserido. Os elementos de uma mesma *thread* são representados por sua distância para esse vídeo-referência comum. A motivação para essa estratégia é a de que os vídeos legítimos, potencialmente, compartilham elementos com o vídeo-referência enquanto que os vídeos *spam* não o fazem, e que isso deva ser verdadeiro independente da *thread* avaliada. Com a estratégia adotada, além de possibilitar a identificação dos vídeos do tipo *spam* em um dado contexto, é possível unificar o treinamento aproveitando o conhecimento aprendido acerca do comportamento dos vídeos legítimos e *spams* para diferentes *threads*.

Na Seção 6.1 é apresentada a metodologia proposta para representação de características visuais sensível ao contexto e nas Subseções 6.2 e 6.3 detalha-se o projeto experimental utilizado em sua avaliação. Os resultados obtidos na avaliação da metodologia são apresentados no Capítulo 7.

6.1 Representação de Características Visuais Sensível ao Contexto

Dadas as características dos vídeos do tipo *spam*, é necessário considerar qual a *thread* em que o vídeo está inserido. Com esse pensamento foi elaborada uma abordagem sensível ao contexto.

A metodologia definida é apresentada na Figura 6.1. As etapas de Pré-processamento e Extração de Características dessa abordagem possuem as mesmas características já apresentadas na descrição da solução modelo no Capítulo 4. Ao final dessas etapas, têm-se o vídeo representado por um vetor de características construído conforme a abordagem tradicional de *bag-of-visual-features*.

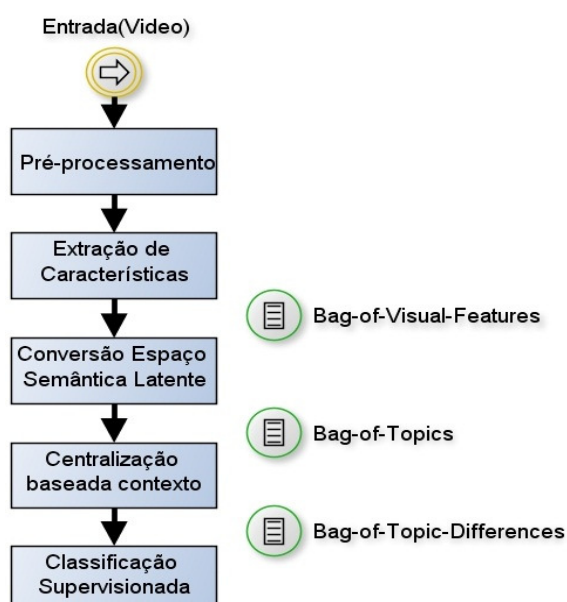


Figura 6.1. Arquitetura da metodologia proposta, com características de ser sensível ao contexto, utilizando a localização contextual dos vídeos, e utilizar o espaço de semântica latente para projeção das características visuais.

Esse vetor de características (*bag-of-visual-features*) é submetido ao algoritmo de análise de semântica latente (LSA), baseado na decomposição de valores singulares (SVD). Então, é efetuada a transformação para um novo espaço de projeção da informação, denominado espaço semântico latente. Nesse processo são identificadas as dimensões mais representativas do conjunto de dados, eliminando as eventuais correlações existentes. O procedimento é basicamente uma transformação de dados, convertendo dados para um novo sistema de coordenadas em que informações redundantes nos atributos são reunidas. O novo vetor de características é denominado *bag-of-topics*.

Em outras palavras, enquanto o *bag-of-visual-features* representa as relações explícitas entre as palavras visuais presentes no vídeo e o dicionário visual utilizado, o *bag-of-topics* consegue mapear padrões ocultos que podem indicar que diferentes palavras visuais estão relacionadas ou, ainda, que certas combinações de palavras visuais fazem sentido quando aplicadas em conjunto.

Após a conversão de espaço, é procedida com a verificação do contexto dos vídeos. Para tal, é medida a similaridade entre cada vídeo e o vídeo-referência da *thread* na qual está inserido. A similaridade é verificada utilizando a distância em cada dimensão de modo independente. Cada vídeo passa então a ser representado pelo vetor de sua distância para o vídeo-referência de sua *thread*. Normalmente, os vetores de características de todos os vídeos utilizam um mesmo ponto como referência para definir a sua escala de valores. O procedimento de verificação de contexto adotado é similar a cada vídeo passar a ter os valores de seu vetor de características com uma escala definida com base em um referencial dependente da *thread* na qual está inserido. A este procedimento é dado o nome de “centralização baseada na localização contextual”.

Antes da verificação do contexto do vídeo, o vetor de características representava informações em um espaço semântico livre de contexto. Agora, o vetor de características é um vetor de diferenças entre cada vídeo e o vídeo-referência de sua *thread*, definindo um espaço contextual. Esta nova situação possibilita trabalhar com diferentes *threads* em um mesmo conjunto de dados e ao mesmo tempo manter um modelo de classificação extremamente simples. Esse novo vetor de características foi denominado de *bag-of-topic-differences*.

Sem o uso do *bag-of-topic-differences* é difícil estabelecer uma comparação entre os vídeos de modo a separar legítimos de *spams*. Com o uso da nossa abordagem, com base na hipótese de que vídeos legítimos possuem conteúdo similar ao do vídeo-referência, vídeos legítimos tendem a possuir um vetor de características com valores mais próximos ao do vídeo-referência de sua *thread*, enquanto que vídeos do tipo *spam* tendem a apresentar valores menos próximos.

6.2 Avaliação de Centralização: Centralização Comum × Centralização baseada na localização contextual

Cada *thread* de discussão pode estar associada a conceitos semânticos completamente diferentes das demais, isso torna difícil estabelecer um critério genérico que separe vídeos legítimos de *spams*. No contexto de redes sociais para compartilhamento de vídeo esse problema é agravado, pois é possível que um mesmo vídeo seja associado a diferentes *threads*, sendo legítimo em um contexto e *spam* no outro.

Com a geração do vetor de características sensível ao contexto esse problema pode ser tratado. Com essa abordagem, até um mesmo vídeo que esteja em *threads* distintas, em uma como legítimo e em outra como *spam*, será representado de modo diferente em cada contexto, pois serão considerados os vídeos-referências de cada uma das *threads* na construção dos vetores de características.

A representação sensível ao contexto também possibilita a manipulação conjunta de diferentes *threads* (possivelmente sobre diferentes temas) mantendo o modelo de classificação extremamente simples (apenas duas classes, legítimo e *spam*, para todo o conjunto de dados) e reduzindo o problema da escassez de elementos no conjunto de treinamento.

A representação sensível ao contexto também possibilita o reuso do conhecimento adquirido acerca dos elementos legítimo e *spam* em diferentes *threads*. Com uma representação relativa ao referencial, o classificador, essencialmente, aprende quais regiões do espaço geométrico são “próximas o suficiente” ao ponto-de-referência, de modo que um elemento presente nessa região seja considerado como “legítimo”. Uma

vez que esse modelo esteja aprendido, novas *threads*, totalmente desconhecidas durante a fase de treinamento, poderão ser classificadas.

A Figura 6.2 apresenta visualmente o contraste entre a medida de similaridade entre vetores de características, tradicional e sensível ao contexto. Na Figura 6.2a é apresentada a abordagem tradicional. Já na Figura 6.2b é apresentado o comportamento utilizando o modelo sensível ao contexto. Para exemplificar, são propostas duas *threads*, representadas, respectivamente, por círculos azuis e triângulos vermelhos. Cada uma possui um elemento representando um vídeo-original, identificados pelas letras **A** e **B**. O elemento indicado pela letra **C** é sugerido como pertencente a ambas as *threads*. Contudo, a distância deste elemento para o vídeo-original de cada uma das *threads* sugere que ele seja legítimo em **A** e *spam* em **B**.

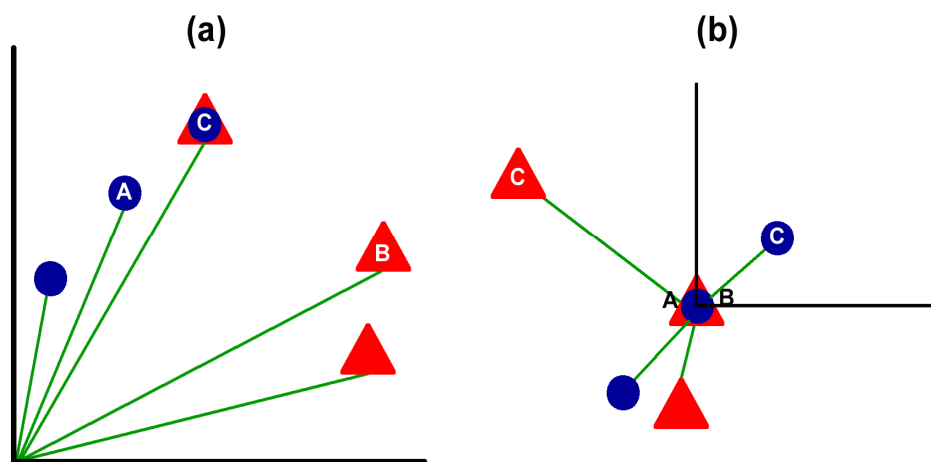


Figura 6.2. Esquema visual contrastando o vetor de características livre de contexto, baseado no *bag-of-visual-features* (a), e o vetor de características sensível ao contexto, baseado no *bag-of-topic-differences* (b).

Na Figura 6.2a o vetor de características dos elementos é descrito com base em um ponto de referência comum para todas as *threads*, neste caso o ponto (0,0) do espaço de características utilizado. Com isso, o elemento representado pela letra **C** possui um vetor de características com os mesmos valores, independente de a quais *threads* seja relacionado. Desse modo, caso seja identificado com classes diferentes na *thread A* e **B** pode induzir o classificador a um erro de classificação. Já na Figura 6.2b, essa situação é contornada pelo uso de um ponto de referência dependente da *thread*. Com isso, o

elemento **C** passa a possuir vetores de características com valores diferentes em cada uma das *threads* a que estiver relacionado.

Neste experimento são avaliados os resultados do uso da centralização baseada na localização contextual e da abordagem tradicional.

6.3 Avaliação de Adequação da Geometria: Espaço de Características × Espaço de Semântica Latente

O uso da representação sensível ao contexto e o fato de as características determinantes na classificação de um vídeo como legítimo ou *spam* serem muito sutis, demandam a necessidade da projeção do vetor de características em um espaço geométrico mais uniforme que o tradicional. A representação dos vídeos através de suas diferenças aos vídeos-referência implica a hipótese dessas diferenças serem comensuráveis ao longo de todo o espaço de representação. Isso, entretanto, pode não ser verdade, à medida que certas regiões do espaço podem ser mais densamente povoadas que outras, tornando a noção de “proximidade” e “vizinhança” dependente da região. Além dessa questão da variação da norma, o “formato” do espaço pode variar consideravelmente de um ponto ao outro, devido à correlações complexas entre as dimensões do vetor de características.

Assim, é interessante para a presente representação adequar o espaço, tornando-o o mais isotrópico possível. A expectativa é fazer com que vídeos legítimos tenham relações semelhantes com os seus vídeos-referência, independentemente da região do espaço onde estes estejam. Além disso, potencialmente, pode haver uma redução na dimensionalidade do vetor de características, o que também ajuda a tornar o espaço mais “bem comportado”.

É importante observar que a adequação do espaço geométrico ocorre antes da etapa de centralização baseada na localização contextual. Desse modo, são preservados os eventuais relacionamentos, explícitos ou implícitos, entre as dimensões do vetor de características. A inversão na ordem dessas etapas pode gerar a perda dessas informações, que são extremamente relevantes para a caracterização do comportamento dos elementos nas classes legítimo e *spam*. Com a adequação do espaço, esses

relacionamentos são identificados e considerados na nova representação do vetor de características, denominada de *bag-of-topics*.

A técnica de *bag-of-visual-features* já é tradicional na literatura da área como a abordagem mais indicada para representar em nível-intermediário as informações extraídas de fontes de baixo-nível de descrição semântica, quer seja em aplicações envolvendo apenas evidências textuais, como utilizado recentemente por Ferragina [2011], ou por Hu et al. [2010] que aplicam na representação de conteúdo visual extraídos de vídeos digitais. Do mesmo modo, o uso de técnicas que projetam o vetor de características em um espaço mais bem comportado e extraem informações latentes contidas no conjunto de dados, tem sido presença recorrente em trabalhos da área, tal como observado recentemente por Zhou et al. [2011], que aplicam o LSA para auxiliar na representação do conteúdo visual no contexto da recuperação de imagens na *web*.

No contexto deste trabalho, é utilizado o LSA para realizar a adequação do espaço de projeção do vetor de características, visto ser uma abordagem tradicional e atender às necessidades inicialmente observadas. Neste experimento são avaliados os resultados obtidos com o uso do *bag-of-visual-features* e *bag-of-visual-differences* contrastando com o uso do *bag-of-topics* e *bag-of-topics-differences*.

Capítulo 7

Abordagem sensível ao contexto – Experimentos e Resultados

Neste capítulo são apresentados a descrição das bases de dados utilizadas e os experimentos realizados para avaliar a metodologia descrita no Capítulo 6. Assim como nos experimentos apresentados no Capítulo 5, neste conjunto de experimentos se enfrenta a dificuldade inerente do escopo escolhido para aplicação deste trabalho, que é a falta de uma referência para comparação dos resultados obtidos. Devido a isso, também foram realizados experimentos com técnicas comuns nas áreas correlatas para servirem de referência. Como nos experimentos executados no Capítulo 5 a solução modelo descrita no Capítulo 4 se mostrou eficaz para o problema de identificação de vídeos do tipo *stuffing*, neste capítulo os experimentos contemplam apenas a tarefa de classificação de vídeos não colaborativos do tipo *spam*.

7.1 Aquisição da Base de Dados

Além da base de dados denominada **selvagem**, descrita na Subseção 5.1.1, o conjunto de experimentos descritos neste capítulo, utilizou uma base de dados sintética, denominada **controlada**. Os experimentos nessas duas bases de dados se apresentam como sendo complementares, uma vez que a base controlada nos permite maior conhecimento e domínio dos contextos dos conteúdos dos vídeos e controle da variabilidade intra e interthreads, possibilitando uma avaliação do poder de generalização e reuso do conhecimento aprendido. Assim, é possível compreender melhor os fenômenos que, por ventura, possam ocorrer e elaborar uma metodologia capaz de lidar com esses fatores.

Por outro lado, a base selvagem oferece uma noção mais clara do comportamento de nossa abordagem em cenários reais.

7.1.1 Base de dados controlada

A base de dados controlada foi construída com base nos resultados de consultas textuais realizadas na rede social YouTube. Foram definidos temas semânticos para cada um dos quais foram construídas consultas por vídeos que possuíssem a informação previamente definida. Os vídeos retornados pelas consultas foram anotados manualmente. Com base nos vídeos legítimos resultantes dessas consultas foram construídas *threads* sintéticas de vídeos-resposta. Cada *thread* foi composta por um número igual de vídeos legítimos e *spam*. Os vídeos do tipo *spam* foram identificados a partir de vídeos legítimos extraídos de outras consultas e que não constaram do conjunto de legítimos de nenhuma outra *thread*.

A Tabela 7.1 apresenta todos os temas, conceitos e palavras-chave utilizadas para construção da base de dados controlada. Foram 10 os temas de alto nível semântico utilizados: Alimentos, Animais, Desenhos, Esportes, Eventos, Filmes, Games, Música, Personalidades e Séries. Para cada tema foram geradas 10 buscas textuais no YouTube. Como exemplo de busca textual utilizada, foram usadas as palavras-chave “steaming” e “asparagus” para buscar por vídeos que atendessem ao conceito de “vídeos ensinando a cozinhar aspargo no vapor”. Nesse caso, foram considerados como vídeos legítimos apenas vídeos que ensinassem a cozinhar aspargo no vapor, não sendo aceitos vídeos com qualquer outro modo de cozimento ou mesmo que ensinassem a cozinhar qualquer outro vegetal no vapor. Foram anotados os 20 primeiros vídeos retornados pela busca. O primeiro vídeo legítimo retornado foi definido como sendo o vídeo original da *thread*, essa postura foi adotada pois o resultado da consulta é ordenado segundo critérios de relevância definidos pela comunidade e, nesse caso, o primeiro colocado seria o mais indicado a atender a consulta textual formulada. Os outros próximos 10 vídeos legítimos foram utilizados como sendo legítimos na *thread* construída, seguindo ideia similar a seleção do primeiro vídeo como sendo o original, os primeiros legítimos são os tidos como mais relevantes pela própria comunidade para atender a consulta formulada. E os demais vídeos identificados como legítimos foram inseridos em outras *threads* na

situação de *spam*, seguindo a ideia de que o *spam* poderia ser um vídeo legítimo se analisado em outro contexto .

Em resumo, a base controlada é composta por 100 *threads* sintéticas, cada uma possuindo 20 vídeos igualmente distribuídos entre legítimos e *spam*. O balanceamento entre os elementos de cada classe nas *threads* foi adotado para minimizar privilegiar a detecção de elementos de uma determinada categoria. Contudo, deve ser ressaltado que essa não é uma situação trivial de ser verificada em ambiente real, podendo ser encontradas desde *threads* que possuem apenas vídeos legítimos até aquelas que possuem apenas vídeos *spam*.

Tabela 7.1. Detalhamento de Temas, Conceitos e Palavras-chave utilizadas na construção da base de dados controlada.

Temas	ID	Conceitos	Palavras-Chave
		Descrição	
Alimentos	Aspargos	Vídeos apresentando o processo de cozimento de aspargos no vapor.	steaming asparagus
Alimentos	Banana	Vídeos apresentando a preparação de torta de banana, sendo necessária a apresentação da receita.	banana pudding recipe
Alimentos	Brócoli	Vídeos apresentando o processo de cozimento de brócolis no vapor.	steaming broccoli
Alimentos	Chocolate	Vídeos apresentando a preparação do bolo de chocolate, mas tem que ser do tipo “lava” e apresentar a receita.	chocolate lava cake recipe
Alimentos	Filé	Vídeos apresentando a preparação de filé mignon grelhado, sendo necessária a apresentação da receita.	grilled filet mignon recipe
Alimentos	Frango	Vídeos apresentando a preparação de frango frito, sendo necessária a apresentação da receita.	fried chicken recipe
Alimentos	Maçã	Vídeos apresentando a preparação de torta de maçã do tipo <i>tatin</i> , sendo necessária a apresentação da receita.	apple tarte tatin recipe
Alimentos	Macarrão	Vídeos apresentando a preparação de macarrão do tipo espaguete a bolonhesa, sendo necessária a apresentação da receita.	spaghetti bolognese recipe
Alimentos	Peru	Vídeos apresentando a preparação de peru assado, sendo necessária a apresentação da receita.	roasted turkey recipe
Alimentos	Salmão	Vídeos apresentando a preparação de salmão grelhado, sendo necessária a apresentação da receita.	grilled salmon recipe
Animais	Cachorro	Vídeos que apresentam o treinamento de cachorros para situações de ataque.	dog attack training video
Animais	Cavalo	Vídeos que apresentam o treinamento de cavalos para situações de salto.	horse training jump videos
Animais	Cobra	Vídeos que apresentam o ataque de cobras a seres humanos.	snake attack human

Animais	Gato	Vídeos engraçados de gatos brincando com caixas.	cat box funny video
Animais	Jacaré	Vídeos que apresentam o ataque de crocodilos a seres humanos.	crocodile attack human
Animais	Leão	Vídeos que apresentam cenas de leões brigando entre si em ambientes selvagens.	lions fighting in the wild
Animais	Macaco	Vídeos que apresentam cenas de macacos brincando entre si.	monkey playing with himself
Animais	Papagaio	Vídeos que apresentam cenas do papagaio Einstein falando.	einstein the talking parrot video
Animais	Peixe	Vídeos que apresentam cenas de pescaria de trutas em rios.	fly fishing trout in river
Animais	Urso	Vídeos que apresentam cenas de ursos polares atacando seres humanos	polar bear attack human
Desenhos	Bob Esponja	Vídeos com cenas do Bob Esponja e Lula Molusco na lanchonete Siri Cascudo, sendo necessário que os dois apareçam juntos na lanchonete em algum momento.	spongebob and squidward at krusty krab
Desenhos	Dragon Ball Z	Vídeos com cenas da luta entre Majin Buu e Goku.	dragon ball z goku vs majin buu fight
Desenhos	Megamente	Vídeos com cenas do Megamente e Roxanne juntos em qualquer situação.	megamind and roxanne together
Desenhos	Meu Malvado Favorito	Vídeos com cenas em que apareçam os Mini-Ons em qualquer situação.	despicable me minion scenes
Desenhos	Muppets	Vídeos com cenas da Miss Piggy e Kermit juntos em qualquer situação.	muppets miss piggy and kermit together
Desenhos	Naruto	Vídeos com cenas da luta entre Pain e Naruto.	naruto vs pain fight scene
Desenhos	Os Sem Floresta	Vídeos com cenas do guaxinin RJ em qualquer situação.	over the hedge rj
Desenhos	Pato Donald	Vídeos com cenas do Pato Donald e dos esquilos Tico e Teco juntos em qualquer situação.	donald duck and chip and dale scenes
Desenhos	Shrek	Vídeos com cenas do Shrek e do burro juntos em qualquer situação.	shrek and donkey scene
Desenhos	Simpsons	Vídeos com cenas internas do bar do Moe em qualquer situação.	moe's bar simpsons
Esportes	Basebol	Vídeos com cenas demonstrando o mecanismo de rebatida, podendo ser em treinamento ou jogos oficiais.	baseball hitting mechanics
Esportes	Basquete	Vídeos contendo cenas de bloqueios (tocos) em partidas de basquetebol, podendo ter outras cenas desde que exiba o instante do bloqueio em algum momento.	basketball match blocks
Esportes	Esqui	Vídeos contendo cenas de acidentes em disputa de esqui, podendo ter outras cenas desde que exiba o instante do acidente em algum momento.	ski crashes scenes
Esportes	Fórmula 1	Vídeos contendo cenas de corridas de Fórmula 1, devendo ser exibidos os carros em algum momento.	formula 1 cars racing

Esportes	Futebol	Vídeos contendo cenas de gol em partidas de futebol, podendo ter outras cenas desde que exiba o instante do gol em algum momento.	soccer match goal
Esportes	Futebol Americano	Vídeos contendo cenas de comemoração após <i>touchdown</i> , podendo ter outras cenas desde que exiba a comemoração em algum momento.	american football touchdown celebrations
Esportes	Futsal	Vídeos contendo cenas de gols em partidas de futsal, podendo ter outras cenas de que exiba o instante do gol em algum momento.	futsal match goals
Esportes	Handebol	Vídeos contendo cenas de gols em partidas de handbol, podendo ter outras cenas de que exiba o instante do gol em algum momento.	handball match goals
Esportes	Motocross	Vídeos contendo cenas de acidentes em corridas de motocross, podendo ter outras cenas de que exiba o instante do acidente em algum momento.	motocross racing crashes
Esportes	Volleybol	Vídeos contendo cenas de treinamento de técnicas de bloqueio em voleybol, podendo ser aceitas cenas de jogos, desde que acompanhadas de explicação dos movimentos.	volleyball blocks training
Eventos	Eleições Americanas	Vídeos contendo cenas em que sejam apresentados os resultados das eleições americanas de 2008, podendo ter outras cenas desde que sejam exibidos os resultados em algum momento.	us election 2008 results
Eventos	Furacão Katrina	Vídeos contendo cenas de resgate após o furacão Katrina.	hurricane katrina rescue footage
Eventos	Greve de Trabalhadores Franceses	Vídeos contendo cenas das paralizações dos trabalhadores franceses em 2010, tendo que ser exibidos o agrupamento de pessoas e faixas/cartazes.	french workers strikes 2010
Eventos	Guerra do Iraque	Vídeos contendo cenas dos soldados americanos em ação na guerra do Iraque, não sendo aceitos vídeos com soldados nos quartéis ou momentos de descanso.	us soldiers action iraq war
Eventos	Mineiros Chilenos	Vídeos contendo cenas do resgate dos mineiros chilenos, sendo necessária a exibição dos mineiros em algum momento.	chilean miners rescue footage
Eventos	Natal	Vídeos contendo imagens de ambientes decorados com luzes de natal, sendo necessária a exibição da iluminação acesa.	christmas lights show
Eventos	Olimpíadas de 2008	Vídeos contendo cenas da cerimônia de abertura dos jogos olímpicos de 2008.	olympic games 2008 opening ceremony
Eventos	Terremoto no Haiti	Vídeos contendo cenas do resgate a população após o terremoto do Haiti.	haiti earthquake rescue videos
Eventos	Tsunami no Japão	Vídeos contendo cenas do resgate a população após o tsunami do Japão.	tsunami japan rescue footage

Eventos	WTC	Vídeos contendo cenas dos aviões colidindo com o WTC.	september 11 wtc plane collision
Filmes	Avatar	Vídeos contendo cenas de Neytiri e Jake juntos.	avatar neytiri and jake scenes
Filmes	Crepúsculo	Vídeos contendo cenas de luta entre vampiros e lobos.	twilight vampire vs werewolf fight scene
Filmes	Harry Potter	Vídeos contendo cenas da batalha entre Harry e Voldemort.	harry potter vs voldemort battle scenes
Filmes	Homem Aranha	Vídeos contendo cenas do instante de transformação com o traje negro (venom), independente do personagem que o utilize.	spiderman 3 movie black suit transformation
Filmes	Piratas do Caribe	Vídeos contendo cenas engraçadas em que apareça o Jack Sparrow.	pirates of the caribbean jack sparrow funny scenes
Filmes	Rocky	Vídeos com cenas de preparação do Rocky para suas lutas, independente de qual edição do filme.	rocky balboa training scenes
Filmes	Thor	Vídeos contendo cenas do Thor e Jane Foster juntos em qualquer situação.	thor movie thor and jane foster
Filmes	Titanic	Vídeos contendo cenas do Jack e Rose juntos em qualquer situação.	titanic jack and rose together
Filmes	Transformers	Vídeos contendo cenas da transformação do bumblebee.	transformers movie bumblebee scene
Filmes	Tron	Vídeos contendo cenas da disputa entre Sam e Rinzler.	tron legacy sam vs rinzler
Games	Angry Birds	Vídeos contendo cenas do jogo Angry Birds na versão easter golden eggs.	angry birds easter golden eggs
Games	Battlefield	Vídeos contendo cenas de disputas <i>online</i> do jogo Battlefield 3.	battlefield 3 online gameplay
Games	Call of Duty	Vídeos contendo cenas do jogo Call of Duty na versão Modern Warfare 3.	call of duty mw3 gameplay
Games	Game of Thrones	Vídeos contendo cenas do jogo "Game of Thrones" na versão Genesis.	game of thrones genesis gameplay
Games	GTA	Vídeos contendo cenas do jogo GTA 5.	gta 5 gameplay
Games	Guittar Hero	Vídeos contendo cenas do jogo Guittar Hero na versão warrior of rock.	guitar hero warriors of rock gameplay
Games	Minercraft	Vídeos contendo cenas do jogo Minercraft na versão survival island.	minecraft survival island gameplay
Games	Mortal Kombat	Vídeos contendo cenas de finalizações no jogo Mortal Kombat 9.	mortal kombat 9 fatalities
Games	Super Mario	Vídeos contendo cenas de disputa com o último chefe do jogo Super Mario Bross.	super mario bros 3 final boss and ending
Games	Warcraft	Vídeos contendo cenas do jogo Warcraft na versão Cataclysm.	warcraft cataclysm gameplay
Musica	Adele	Vídeos contendo cenas de shows ao vivo de Adele, sendo necessária a exibição de cenas do artista e do público em algum momento.	adele live show

Musica	Bruno Mars	Vídeos contendo cenas de shows ao vivo de Bruno Mars, sendo necessária a exibição de cenas do artista e do público em algum momento.	bruno mars live show
Musica	Cold Play	Vídeos contendo cenas de shows ao vivo de Cold Play, sendo necessária a exibição de cenas do artista e do público em algum momento.	cold play live show
Musica	Jennifer Lopez	Vídeos contendo cenas de shows ao vivo de Jennifer Lopez, sendo necessária a exibição de cenas do artista e do público em algum momento.	jennifer lopez live show
Musica	Justin Bieber	Vídeos contendo cenas de shows ao vivo de Justin Bieber, sendo necessária a exibição de cenas do artista e do público em algum momento.	justin bieber live show
Musica	Katty Perry	Vídeos contendo cenas de shows ao vivo de Katty Perry, sendo necessária a exibição de cenas do artista e do público em algum momento.	katy perry live show
Musica	Mariah Carey	Vídeos contendo cenas de shows ao vivo de Mariah Carey, sendo necessária a exibição de cenas do artista e do público em algum momento.	mariah carey live show
Musica	Maroon 5	Vídeos contendo cenas de shows ao vivo de Maroon 5, sendo necessária a exibição de cenas do artista e do público em algum momento.	maroon 5 live show
Musica	Michael Jackson	Vídeos contendo cenas de Michael Jackson dançando.	michael jackson dancing videos
Musica	Shakira	Vídeos contendo cenas de Shakira dançando em show ao vivo, sendo necessária a exibição de cenas do artista e do público em algum momento.	shakira dancing live show videos
Personalidades	Barack Obama	Vídeos contendo cenas do discurso de Barack Obama após a vitória nas eleições de 2008.	barack obama victory speech 2008
Personalidades	Bill Clinton	Vídeos contendo cenas de entrevista com Bill Clinton em qualquer momento da história.	bill clinton interview
Personalidades	George Bush	Vídeos contendo cenas de visitas de George Bush ao Iraque durante a Guerra.	bush visits iraq
Personalidades	Jô Soares	Vídeos contendo cenas do Jô Soares fazendo entrevistas em seu programa.	jo soares show interview
Personalidades	Omar Kadaff	Vídeos contendo cenas da captura de Omar Kadaff.	gaddafi captured footage
Personalidades	Oprah Winfrey	Vídeos contendo cenas da Oprah Winfrey fazendo entrevistas em seu programa.	oprah winfrey show interview
Personalidades	Osama Bin Laden	Vídeos contendo cenas de entrevistas de Osama Bin Laden em qualquer momento da história.	osama bin laden interview
Personalidades	Rainha Elizabeth	Vídeos contendo cenas da rainha Elizabeth 2 acompanhada de seu esposo príncipe Philip em qualquer situação.	queen elizabeth 2 and prince philip

Personalidades	Saddan Russein	Vídeo contendo cenas da captura de Saddam Russein.	saddam hussein captured footage
Personalidades	William e Kate	Vídeos contendo cenas do príncipe William junto com sua esposa Kate durante a cerimônia de casamento, sendo aceitos vídeos de todos os eventos do casamento.	william and kate together wedding ceremony
Séries	Board Walk Empires	Vídeos contendo cenas com o Nucky em qualquer situação.	boardwalk empire nucky scenes
Séries	CSI NY	Vídeos contendo cenas de beijo entre Danny e Lindsay em qualquer situação.	csi new york danny and lindsay kiss scenes
Séries	Game of Thrones	Vídeos contendo cenas de lutas entre Lannister e os Stark em qualquer situação.	game of thrones stark vs lannister
Séries	House	Vídeos contendo cenas engraçadas envolvendo o personagem do Dr. Wilson.	house md wilson funny scenes
Séries	Law e Order SVU	Vídeos contendo cenas em que apareçam juntos os personagens Stabler e Benson em qualquer situação.	law and order svu benson and stabler together scenes
Séries	Monk	Vídeos contendo cenas engraçadas envolvendo os personagens do Monk e da Natalie.	monk and natalie funny scene
Séries	NCIS	Vídeos contendo cenas em que apareçam juntos os personagens Gibbs e Abby em qualquer situação.	ncis gibbs and abby scenes
Séries	Psych	Vídeos contendo cenas engraçadas envolvendo os personagens do Shawn e Gus.	psych shawn and gus funny scenes
Séries	The Good Wife	Vídeos contendo cenas em que apareçam juntos os personagens Alicia e Will em qualquer situação.	the good wife alicia and will scenes
Séries	Two and a Half Man	Vídeos contendo cenas da participação do Ashton Kutcher em qualquer situação.	two and a half men ashton kutcher scene

7.2 Pseudoalgoritmo para abordagem sensível ao contexto

Na Tabela 7.2 é apresentado o algoritmo elaborado para a execução da metodologia sensível ao contexto descrita no Capítulo 6. A arquitetura apresentada na Figura 6.1 é utilizada como referência para definição das diferentes fases da metodologia. A etapa de Pré-processamento (linhas 1 e 2) é dependente da característica visual a ser extraída, por exemplo: no caso do SIFT, histogramas e outros que trabalham com base em imagens, são executados a divisão do vídeo em quadros no formato *.jpg*, com taxa de 01

quadro/segundo, para então ser gerado o resumo automático do vídeo utilizando o algoritmo VSUMM [Avila et al., 2008a]; para o STIP o vídeo de entrada é convertido para o formato .avi, caso não esteja nesse formato.

Tabela 7.2. Algoritmo utilizado para execução da abordagem sensível ao contexto.

Entrada:	
$V[n]$: representa o conjunto de vídeos, sendo n o número de vídeos presente no conjunto (incluindo vídeos de treinamento, testes e originais).	
Saída:	
Um conjunto de vídeos associados a suas respectivas classes.	
1.	for v in V do
2.	$Vp[i]$ = prepara (conversão de formato, amostragem, resumo estático) o vídeo v ;
3.	for vp in Vp do
4.	$Vf[i]$ = características visuais (SIFT, STIP ou outra) de vp ;
5.	VOC = dicionário visual obtido através do total de características em Vf (todos os vídeos), contendo w palavras visuais;
6.	for vf in Vf do
7.	BOVF[i] = <i>bag-of-visual-features</i> para o vídeo vf com base em VOC. Cada <i>bag</i> é um vetor com w entradas (uma entrada para cada palavra do dicionário). O <i>bag</i> faz uma contagem simples, atribuindo as características em vf à sua palavra visual mais próxima no dicionário;
8.	$TF_IDF_{n \times w}$ = matriz BOVF $_{n \times w}$ normalizada utilizando TF-IDF;
9.	$U_{n \times n}, \Sigma_{n \times w}, V_{w \times w}$ = decomposição de valores singulares em $TF_IDF_{n \times w}$;
10.	$U_{n \times k}$ = os k tópicos mais relevantes em $U_{n \times n}$;
11.	for u in U do
12.	BOTD[i] = diferença absoluta, componente por componente, entre o vetor u e o vetor $U[r]$ (r sendo o vídeo referência de u);
13.	return BOTD[];

A etapa de Extração de Características Visuais (linhas 3 a 7) é dependente da característica visual a ser extraída, por exemplo: no caso do SIFT é adotado algoritmo original de Lowe [1999], utilizando suas configurações padrão; para o STIP é adotado o algoritmo original proposto por Laptev [2005], com descritor HOG-HOF e demais configurações padrão.

Ainda na etapa de Extração de Características Visuais, o vetor de características é representado conforme a técnica de *bag-of-visual-features* (linhas 5 a 7). Para tal, é

construído um dicionário (linha 5), que considera cada vídeo como sendo um conjunto de palavras-visuais (cada região/ponto de interesse descrito no vídeo é entendido como sendo uma palavra-visual) e seleciona, aleatoriamente, um conjunto de w palavras pertencentes a todo o conjunto de vídeos. Neste trabalho foi adotado $w = 5000$. Como restrição, foi definido que cada uma das classes de vídeos (legítimo e *spam*) deveria contribuir com igual número de palavras para evitar um eventual desbalanceamento (viés) na representação do conhecimento.

A etapa de Conversão para o Espaço de Semântica Latente é executada nas linhas 8 a 10. Inicialmente é aplicada a técnica TF-IDF (linha 8), avaliando o quão importante uma palavra visual é para um dado documento no conjunto. Deve ser ressaltado que existem diferentes formas de se calcular o TF-IDF e neste trabalho foi adotado o que segue:

- $TF[v][w]$: frequência do termo w no documento v / frequência de todos os termos no vídeo v , e;
- $IDF[v][w]$: $\log$ (número de vídeos no conjunto / número de documento com a palavra w).

Ainda na mesma etapa, após o TF-IDF, é aplicada a decomposição de valores singulares (linha 9) utilizando a implementação padrão disponibilizada no pacote *scipy* da linguagem *Python*. Posteriormente é realizada a seleção de tópicos relevantes (linha 10), sendo mantidas todas as colunas (tópicos latentes) que possuem valores singulares maiores que zero.

Na etapa de Centralização Baseada no Contexto (linhas 11 e 12) é utilizada a distância euclidiana como métrica para aferir a distância entre um dado vídeo e o seu respectivo vídeo-referência no espaço de projeção adotado. A etapa de Classificação Supervisionada, presente na Figura 6.1, não está representada no algoritmo por ser executada utilizando a implementação de SVM, com *kernel* linear, disponível no pacote *libsvm* (Chang e Lin, 2011).

7.2.1 Análise de Custos

Para utilização da metodologia proposta em ambiente real é importante que seja feita uma análise dos custos envolvidos em sua execução. O custo da etapa de pré-

processamento é dependente de duas etapas distintas: a primeira é o particionamento do vídeo em quadros, ou conversão para outro formato, de acordo com o tipo de característica visual a ser extraída, ambas possuindo um custo $O(v)$, sendo v a quantidade de vídeos no conjunto de entrada, e; a segunda, sendo a geração do resumo do vídeo, executada quando do uso de extratores de características visuais que se baseiam em quadros (globais ou locais puramente espaciais), que possui custo $O(v*q)$, sendo q a quantidade de quadros que compõem o resumo de um dado vídeo.

A extração de características visuais tem seu custo de execução associado à técnica de extração adotada. Cabe ressaltar que as abordagens que se apresentaram como mais eficazes neste trabalho, SIFT e STIP, possuem alto tempo/custo de execução. Mais especificamente no caso do SIFT, a redução na quantidade de quadros a serem processados, efetivada com o resumo gerado para cada vídeo, reduz consideravelmente o tempo de execução, com casos em que chega à ordem de 95%. Cabe ressaltar que nesse caso, a representatividade do conteúdo presente no vídeo fica sob a responsabilidade da abordagem adotada para geração do resumo automático, o que pode acarretar em perda de qualidade na tarefa de classificação.

Uma decisão que afeta consideravelmente o custo de execução da abordagem elaborada, está associada ao número de palavras (w) que irão compor o dicionário visual. Baseado na literatura da área (tais como Jurie e Triggs [2005], Sivic e Zisserman [2008], entre outros) foi observado que $w = 1000, 2500$ e 5000 são adotados em várias ocasiões relatando bons resultados. Ainda, indicam que quanto maior o dicionário, maior o custo de execução, porém, tende a se alcançar melhores resultados. O custo para construção do dicionário também é dependente da abordagem utilizada para identificação das palavras-visuais que irão compor o dicionário, como neste trabalho adotou-se uma abordagem de seleção aleatória, esse custo é linear, associado ao número de palavras-visuais a serem selecionadas. Cabe ressaltar que a definição de valor “ótimo” para a quantidade de palavras-visuais, ou mesmo a própria construção de vocabulários visuais são tarefas em aberto na área.

A quantidade de palavras no dicionário visual afeta diretamente a construção do *bag-of-visual-features*, visto que esse é gerado a partir da mensuração da distância de cada palavra-visual (p) presente em um vídeo (v) para cada uma das palavras-visuais presentes

no dicionário (w), o que acarreta em um custo total $O(v*p*w)$. Cabe ressaltar que as palavras-visuais são, geralmente, vetores (128 dimensões no caso do SIFT ou 162 dimensões no caso do STIP).

O tamanho do dicionário visual também influencia no custo da etapa de conversão do espaço de projeção das características visuais, executados com base no TF-IDF e SVD. Contudo, um fator de relevância neste processo, e que praticamente inviabiliza a adoção da metodologia em ambientes reais, é a execução da decomposição de valores singulares (SVD). A execução “padrão” do SVD exige o conhecimento de todo o conjunto de vídeos (treinamento e testes) para obtenção dos valores singulares. Ou seja, a cada novo vídeo adicionado a ser classificado é necessário um novo cálculo do SVD sobre toda a base de vídeos. Existem algumas alternativas para minimizar os custos com esse procedimento, como a apresentada por Tougas e Spiteri [2008], que prevê o reaproveitamento das matrizes já calculadas para obtenção dos valores singulares de um novo elemento. Contudo, os resultados não são exatamente os mesmos que a abordagem padrão, o que pode influenciar diretamente na qualidade dos resultados da tarefa de classificação, sendo necessário avaliar o compromisso custo/benefício de adoção desse tipo de estratégia.

Os tempos exatos de execução dependem de vários aspectos de implementação e tecnologia, mas para fins ilustrativos, no processamento do experimento *Spam* x Legítimo, com 2.000 vídeos, no total de 91,27 horas, utilizando um equipamento com processador de 8 núcleos de alta-capacidade, foram necessários, aproximadamente: 30 horas na extração do STIP; 41 horas na extração do SIFT; 47 horas na construção do *bag-of-visual-features*; 1,2 horas na conversão para o espaço latente, e; 0,1 horas na centralização baseada no contexto.

Além de custos computacionais, como os descritos acima, o custo humano envolvido na construção da base de dados é considerável e não pode ser deixado de fora. A tarefa de anotação manual do conjunto de treinamento consome muito tempo e é extremamente importante para a garantia de qualidade dos resultados. A definição do tamanho do conjunto de treinamento e estratégia para a execução da anotação desse conjunto são decisões que impactam nesse custo.

7.3 Visão Geral dos Experimentos

Nesta seção são apresentados detalhes dos experimentos realizados para verificar o desempenho da abordagem sensível a contexto, *bag-of-topic-differences* (BoTD), descrita no Capítulo 6. Essa metodologia utiliza uma representação do vetor de características baseada na localização contextual dos vídeos e projeção desse vetor para o espaço latente. Como referências de comparação são avaliados o *bag-of-visual-features* tradicional (BoVF) e outras duas representações, denominadas *bag-of-topics* (BoT), que usa o espaço de tópicos mas não incorpora o contexto, e *bag-of-visual-differences* (BoVD), que incorpora o contexto mas não usa o espaço de tópicos. Esses *baselines* permitem avaliar as contribuições relativas de cada componente (espaço de tópicos, centralização baseada no contexto) da representação proposta.

Como descritores de informação visual de baixo-nível são adotados o SIFT e STIP, motivado pelos resultados dos experimentos apresentados no Capítulo 5. Os experimentos foram realizados utilizando tanto a base de dados controlada quanto a selvagem. Foi adotada a abordagem de validação cruzada com 5 dobras (*5-folds cross-validation*, em inglês).

Para os experimentos que utilizam a base controlada, o conjunto de treinamento e testes foi organizado visando garantir maior poder de generalização aos resultados. Em cada *fold* o conjunto de testes foi formado com os vídeos de dois conceitos semânticos e os vídeos dos demais conceitos formaram o conjunto de treinamento. Por exemplo, quando o conjunto de testes possuiu todas as *threads* dos conceitos Alimentos e Animais, o conjunto de treinamentos foi formado por todas as *threads* dos conceitos Desenhos, Esportes, Eventos, Filmes, Games, Música, Personalidades e Séries. Com isso, foi proporcionada uma situação em que os vídeos legítimos do conjunto de testes não possuem nenhuma relação semântica com os do conjunto de treinamento.

Na base de dados selvagem, cada *fold* foi construído utilizando 400 vídeos para testes e 1.600 para treinamento. Em ambos os conjuntos os elementos foram divididos igualmente entre as duas classes, legítimo e *spam*. O objetivo de refazer os experimentos na base selvagem é demonstrar o desempenho em situações reais.

Em todos os experimentos é utilizado o classificador SVM com *kernel* linear. Os dados de entrada são normalizados antes da execução do classificador, utilizando uma escala $[-1, +1]$. Como instrumentos de observação e de avaliação da qualidade dos resultados obtidos são utilizadas a ANOVA em conjunto com o teste-t, a curva ROC e os valores médios de Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Área Abaixo da Curva (AUC) considerando os 5 *folds*. Em todas as situações, o ANOVA foi executado com base nos valores de AUC. O limite adotado para rejeição de hipótese-nula foi $\alpha = 0,05$ (nível de confiança de 95%). Em todos os casos a hipótese-nula é de que não existe diferença significativa nos resultados avaliados.

Utilizando a metodologia apresentada no Capítulo 6, é procedida a verificação da validade de algumas hipóteses de pesquisa levantadas neste trabalho, conforme segue:

- Uma representação que incorpora o contexto no qual o vídeo está sendo analisado proporciona maior qualidade nos resultados da classificação em comparação ao uso da representação de *bag-of-visual-features* tradicional;
- Uma transformação na projeção do espaço de representação das características de nível-médio melhora a representação sensível ao contexto, tornando as distâncias, entre vídeos relacionados, mais uniformes em todo o espaço.

7.4 Avaliando a adoção da representação baseada na localização contextual dos vídeos e a transformação linear do espaço de projeção em ambiente controlado

O desempenho do uso da representação do vetor de características baseada na localização contextual dos vídeos aliada à projeção desse vetor no espaço latente é avaliado contrastando o resultado dessa abordagem com etapas “intermediárias” desse processo. Na Tabela 7.3 são apresentados os valores médios de Área Abaixo da Curva (AUC) considerando os 5 *folds* de cada experimento. Na Tabela 7.4 são apresentados os resultados do teste de significância par-a-par com confiança de 95%. A ANOVA para esse conjunto de experimentos obteve $p\text{-value} = 0,0006$, indicando a significância da diferença entre as médias de AUC dos experimentos com nível de confiança de 95%.

Conforme observado na Tabela 7.3 os experimentos com maiores valores médios de AUC são os que utilizam a metodologia proposta neste trabalho, BoTD: 0,930 (STIP) e 0,914 (SIFT); e não é possível afirmar que existe a superioridade de um descritor em relação ao outro, com o uso dessa abordagem, pois o teste de significância estatística com nível de confiança de 95% não rejeitou a hipótese-nula de que não existe uma diferença significativa entre os resultados.

Tabela 7.3. Valores médios de Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Área Abaixo da Curva (AUC) nos experimentos *Bag-of-visual-features* (BoVF), *Bag-of-visual-differences* (BoVD), *Bag-of-topics* (BoT) e *Bag-of-topics-differences* (BoTD), utilizando os descritores SIFT e STIP na base controlada.

		TPR	FPR	AUC
BoVF – SIFT	1	0,775	0,230	0,801
BoT – SIFT	2	0,507	0,588	0,441
BoVD – SIFT	3	0,555	0,372	0,589
BoTD – SIFT	4	0,908	0,179	0,914
BoVF – STIP	5	0,822	0,404	0,716
BoT – STIP	6	0,498	0,554	0,416
BoVD – STIP	7	0,590	0,393	0,613
BoTD – STIP	8	0,922	0,152	0,930

Avaliando o desempenho dos referenciais, pode ser observado que a estratégia tradicional de BoVF supera as duas outras abordagens, BoT e BoVD, com ambos descritores, mas somente nos experimentos com o descritor SIFT essa diferença é validada pelo teste de significância estatística, com o STIP apenas a diferença para o BoT é comprovada como significativa. Já a vantagem da abordagem BoVD em comparação ao uso do BoT não é validada na análise par-a-par por não rejeitar a hipótese-nula.

Fazendo a comparação entre as abordagens de BoVF e BoTD, com um mesmo descritor: no caso do SIFT, a análise par-a-par não rejeita a hipótese-nula, não permitindo a inferência de superioridade na adoção da estratégia de BoTD; já com o uso do STIP, o teste par-a-par rejeita a hipótese-nula, o que indica a significância estatística da diferença entre os resultados e possibilita a afirmação de que a abordagem de BoTD apresenta melhor desempenho que a tradicional BoVF.

Tabela 7.4. Matriz com resultados do teste-t par-a-par, com significância de 95%, nos experimentos *Bag-of-visual-features* (BoVF), *Bag-of-visual-differences* (BoVD), *Bag-of-topics* (BoT) e *Bag-of-topics-differences* (BoTD), na base controlada.

		1	2	3	4	5	6	7	8
BoVF – SIFT	1		○	○	×	×	○	+	×
BoT – SIFT	2	○		×	·	+	×	×	·
BoVD – SIFT	3	○	×		●	×	×	×	●
BoTD – SIFT	4	×	·	●		○	●	●	×
BoVF – STIP	5	×	+	×	○		×	×	○
BoT – STIP	6	○	×	×	●	×		×	·
BoVD – STIP	7	+	×	×	●	×	×		●
BoTD – STIP	8	×	·	●	×	○	·	●	

Legenda:

Rejeição de Hipótese Nula

< 0,001

·

< 0,01

●

< 0,05

○

Não Rejeição de Hipótese Nula

< 0,1

+

>= 0,1

×

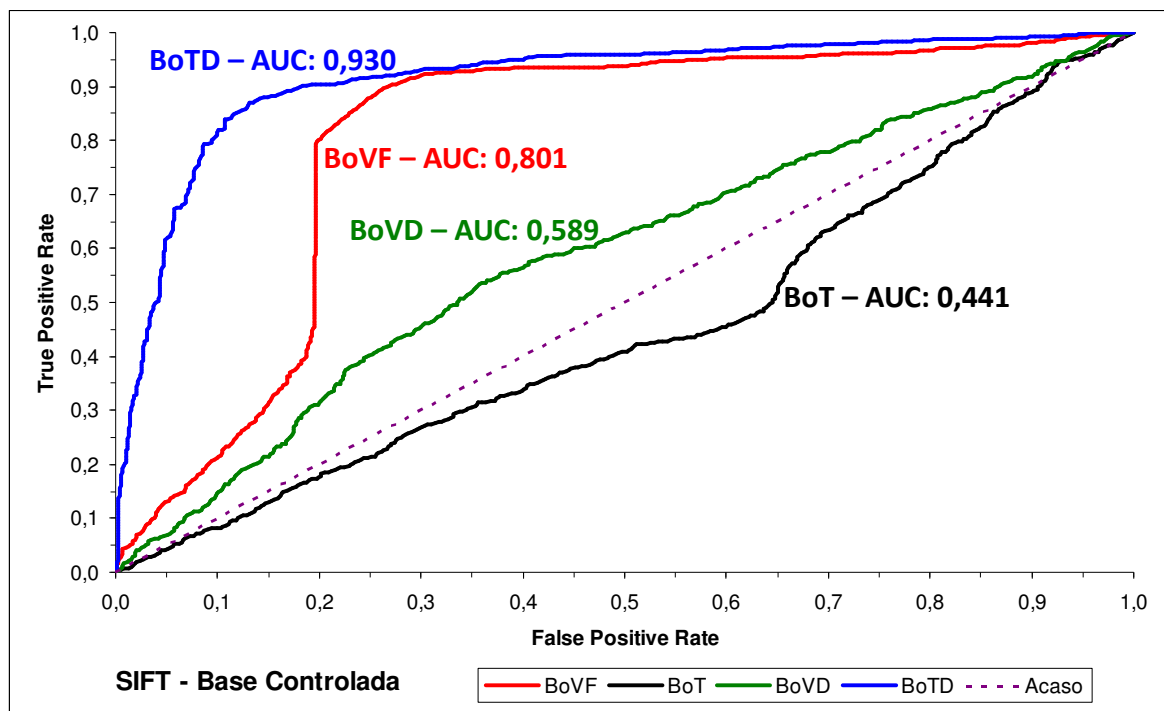


Figura 7.1. Curva ROC para avaliação dos experimentos *Bag-of-visual-features* (BoVF), *Bag-of-visual-differences* (BoVD), *Bag-of-topics* (BoT) e *Bag-of-topics-differences* (BoTD), na base controlada, utilizando o descritor SIFT.

Ainda assim, essa situação desperta a atenção e cabe uma análise das Curvas ROC, apresentadas nas Figuras 7.1 (SIFT) e 7.2 (STIP), e uma reflexão quanto a construção da base de dados controlada.

Em ambas as figuras é observado que a curva BoVF apresenta inicialmente um desempenho inferior ao BoTD e um crescimento vertiginoso próximo do instante $TPR = 0,2$. Esse crescimento faz com que o desempenho dessa abordagem se aproxime ao do BoTD. Faz-se necessário então compreender as causas desse fenômeno.

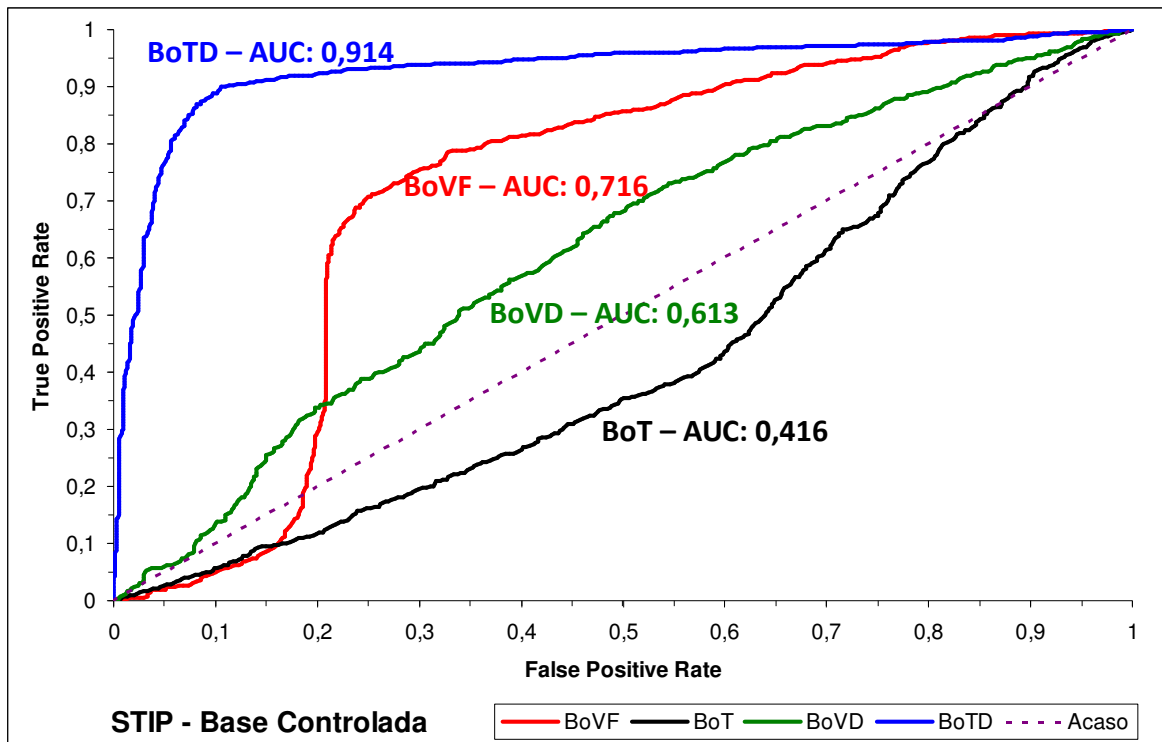


Figura 7.2. Curva ROC para avaliação dos experimentos *Bag-of-visual-features* (BoVF), *Bag-of-visual-differences* (BoVD), *Bag-of-topics* (BoT) e *Bag-of-topics-differences* (BoTD), na base controlada, utilizando o descritor STIP.

A estratégia adotada para construção da base controlada proporcionou ao classificador baseado na abordagem BoVF “aprender” o modelo dos elementos do tipo *spam* impactando diretamente na obtenção de resultados acima do inicialmente esperado. Ocorre que os vídeos utilizados como *spam* foram selecionados aleatoriamente dentre os vídeos anotados como legítimo e que não foram aproveitados como tal em sua *thread*. Essa estratégia desencadeou um evento que contribuiu positivamente com o desempenho da abordagem de BoVF: vídeos *spam* de um mesmo conceito podem ter sido utilizados em *threads* tanto da base de teste quanto na base de treinamento para um determinado *fold*. Entendendo que vídeos de mesmo conceito tendem a possuir elementos visuais similares, uma vez conhecido o vídeo na base de treinamento o classificador possui maiores chances de classificá-lo corretamente.

Essa situação foi tolerada, visto que o mesmo evento não favorece no desempenho da metodologia proposta (BoTD). Como os elementos são representados com base na distância para um referencial comum de acordo com a *thread* em que está inserido, mesmo elementos idênticos inseridos em *threads* diferentes são representados por vetores de características distintos, conforme exemplificado na Seção 6.2.

Um efeito positivo das avaliações na base controlada é a comprovação do poder de generalização do conhecimento que a abordagem de BoTD possui. A representação sensível ao contexto aliada à projeção no espaço latente é capaz de descrever o comportamento de elementos legítimos e *spam* independente do conceito que estejam apresentando. Isso é demonstrado pelos altos índices AUC obtidos em experimentos nos quais o conjunto de treinamento não conhece nenhuma informação dos conceitos a serem classificados. Essa característica possibilita o enfrentamento do problema da escassez de elementos para o conjunto de treinamento em cada *thread*, observado em situações reais de aplicações no ambiente *web*.

7.5 Avaliando a adoção da representação baseada na localização contextual dos vídeos e a transformação linear do espaço de projeção em ambiente real

Os experimentos comparativos entre a metodologia de representação sensível ao contexto e as estratégias adotadas como referência são reavaliados utilizando o conjunto de dados da base selvagem, que representa em menor escala o ambiente real de uma rede social. Na Tabela 7.5 são apresentados os valores médios de Área Abaixo da Curva (AUC) obtidos por cada um dos experimentos. Na Tabela 7.6 são apresentados os resultados do teste de significância par-a-par com confiança de 95%. O *p-value* obtido pela ANOVA nesse conjunto de experimentos é $<0,0001$, o que garante a significância estatística das diferenças de AUC na população com confiança superior a 95%.

Assim, como observado nos experimentos com a base controlada, os melhores resultados foram obtidos com o uso da metodologia proposta: 0,937 (SIFT) e 0,868 (STIP). E, diferentemente do que ocorreu na base controlada, a análise par-a-par rejeitou a hipótese-nula de que as diferenças entre esses resultados não possuem significância. Com

7.5 – AVALIANDO A ADOÇÃO DA REPRESENTAÇÃO BASEADA NA LOCALIZAÇÃO CONTEXTUAL DOS VÍDEOS E A TRANSFORMAÇÃO LINEAR DO ESPAÇO DE PROJEÇÃO EM AMBIENTE REAL 98

isso, é possível afirmar que o SIFT apresenta melhor desempenho que o STIP na realização da tarefa avaliada.

Deve ser ressaltada a robustez demonstrada pelo SIFT, se comparadas as avaliações em ambiente controlado e selvagem o valor de AUC foi melhorado de 0,914 (controlada) para 0,937 (selvagem). Por outro lado, o STIP demonstrou queda de desempenho no ambiente selvagem em contraste com o ambiente controlado.

Tabela 7.5. Valores médios de Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR) e Área Abaixo da Curva (AUC) nos experimentos *Bag-of-visual-features* (BoVF), *Bag-of-visual-differences* (BoVD), *Bag-of-topics* (BoT) e *Bag-of-topics-differences* (BoTD), utilizando os descritores SIFT e STIP na base selvagem.

		TPR	FPR	AUC
BoVF – SIFT	1	0,490	0,339	0,582
BoT – SIFT	2	0,569	0,541	0,596
BoVD – SIFT	3	0,627	0,247	0,718
BoTD – SIFT	4	0,891	0,169	0,937
BoVF – STIP	5	0,679	0,583	0,577
BoT – STIP	6	0,569	0,543	0,592
BoVD – STIP	7	0,689	0,337	0,728
BoTD – STIP	8	0,847	0,257	0,868

Em comparação com as demais técnicas de referência, a metodologia proposta demonstrou-se superior nos experimentos com ambos os descritores e em todos os fatores avaliados. Na análise par-a-par teve rejeitada a hipótese-nula de que não haveria diferença significativa entre os resultados, com nível de confiança de 95%. Essa situação indica claramente a vantagem da adoção da estratégia de BoTD proposta neste trabalho. As Figuras 7.3 e 7.4 apresentam os gráficos com as Curvas ROC geradas para cada um dos experimentos utilizando os descritores SIFT e STIP, respectivamente.

Avaliando as estratégias de referência é observado um comportamento diferente do verificado na base controlada. O BoVD, estratégia que incorpora a representação sensível ao contexto ao BoVF, demonstra melhor desempenho que as demais abordagens que não utilizam a representação baseada na localização contextual dos vídeos. Ainda, é observado que a estratégia de projeção do BoVF no espaço latente, BoT, apresenta melhor desempenho que o BoVF tradicional. Porém, em nenhuma situação a análise de significância estatística rejeitou a hipótese-nula de que não há diferença significativa entre os resultados, com confiança de 95%.

Tabela 7.6. Matriz com resultados do teste-t par-a-par, com significância de 95%, nos experimentos *Bag-of-visual-features* (BoVF), *Bag-of-visual-differences* (BoVD), *Bag-of-topics* (BoT) e *Bag-of-topics-differences* (BoTD), na base selvagem.

		1	2	3	4	5	6	7	8
BoVF – SIFT	1		×	.	.	×	×	.	.
BoT – SIFT	2	×		.	.	+	×	.	.
BoVD – SIFT	3	.	.		.	.	.	×	.
BoTD – SIFT	4	.	.	.		.	.	.	.
BoVF – STIP	5	×	+	.	.		×	.	.
BoT – STIP	6	×	×	.	.	×		.	.
BoVD – STIP	7	.	.	×	.	.	.		.
BoTD – STIP	8	.	.	.	.	.	.	.	

Legenda:

Rejeição de Hipótese Nula

< 0,001 .

< 0,01 •

< 0,05 ○

Não Rejeição de Hipótese Nula

< 0,1 +

>= 0,1 ×

Contudo, essas verificações reforçam a indicação de que o bom desempenho da abordagem tradicional de BoVF na base controlada se deu apenas devido ao modo como as *threads* foram construídas. Em uma situação real, todas as tentativas de aumentar o poder de informação do vetor de características se demonstraram com qualidade no mínimo comparável a da abordagem tradicional.

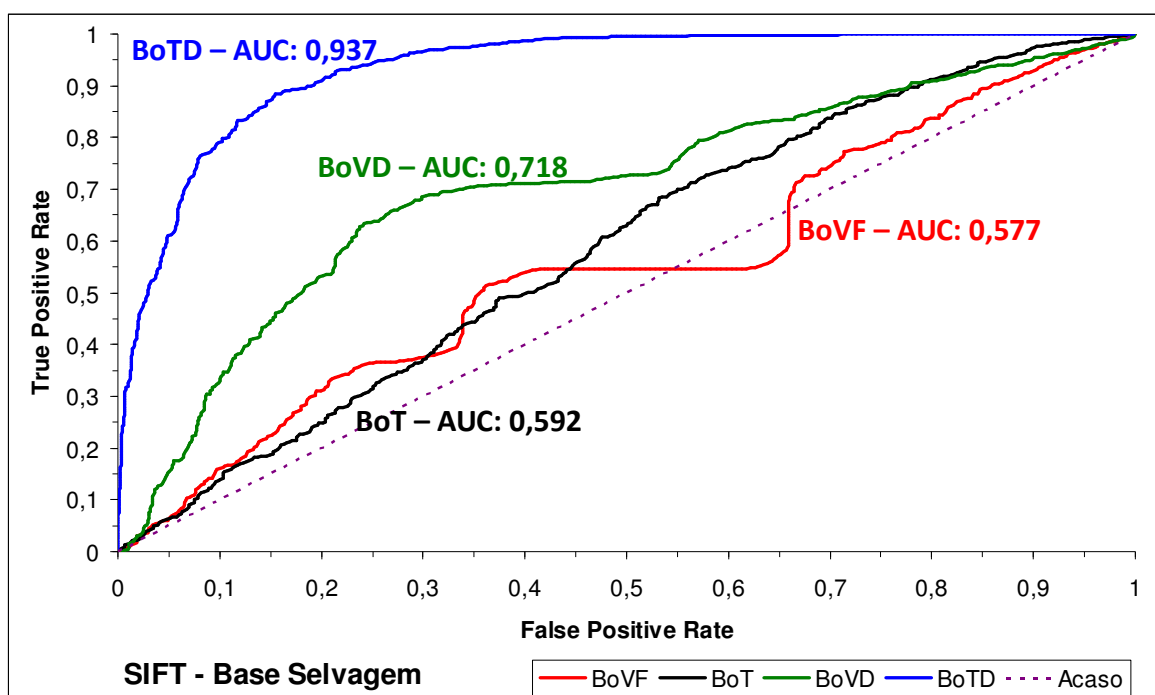


Figura 7.3. Curva ROC para avaliação dos experimentos *Bag-of-visual-features* (BoVF), *Bag-of-visual-differences* (BoVD), *Bag-of-topics* (BoT) e *Bag-of-topics-differences* (BoTD), na base selvagem, utilizando o descritor *SIFT*.

7.6 Considerações

Com a adoção da metodologia proposta foi possível obter considerável ganho de desempenho, demonstrado pelo aumento de aproximadamente 30% no AUC se comparada à abordagem tradicional de BoVF.

Mais especificamente, as avaliações das Hipóteses 4 e 5, realizadas nos experimentos descritos nas Seções 7.3 (base controlada) e 7.4 (base selvagem), demonstram claramente a validade dessas e ressaltam o papel fundamental que possuem na tarefa de distinção entre elementos das classes legítimo e *spam*. Ainda, o uso de um descritor de nível-intermediário que utiliza apenas a representação sensível ao contexto sem a etapa de transformação do espaço de projeção (BoVD), possibilitou a verificação da relevância do uso combinado dessas estratégias, demonstrando que isoladamente, mesmo superando a técnica tradicional de BoVF, a representação sensível ao contexto não atinge todo o seu potencial.

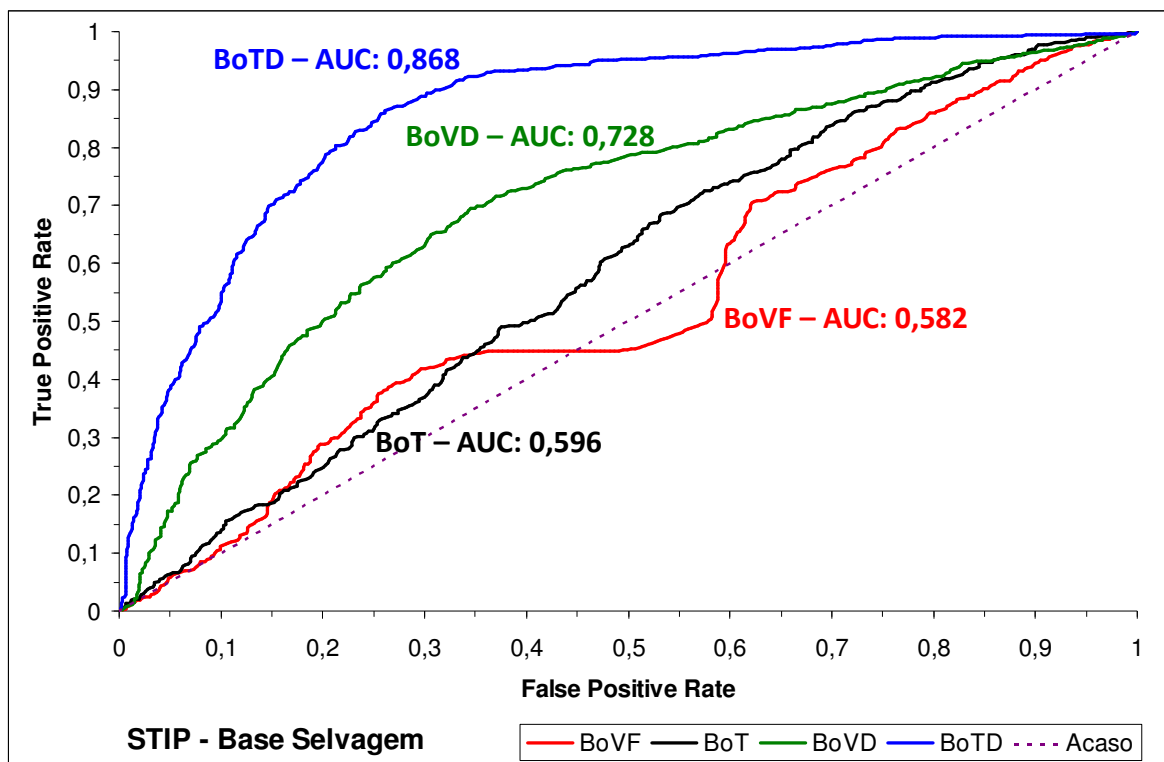


Figura 7.4. Curva ROC para avaliação dos experimentos *Bag-of-visual-features* (BoVF), *Bag-of-visual-differences* (BoVD), *Bag-of-topics* (BoT) e *Bag-of-topics-differences* (BoTD), na base selvagem, utilizando o descritor STIP.

Capítulo 8

Conclusão

Neste capítulo são apresentadas as considerações finais acerca das atividades desenvolvidas neste trabalho e os artigos científicos aprovados com divulgação dos resultados.

8.1 Considerações Finais

O aumento no volume de utilização de redes sociais virtuais para compartilhamento de vídeo tem gerado novos desafios para a computação. Dentre eles, a construção de ferramentas capazes de tornar mais agradável a experiência no uso dessas comunidades por parte de seus membros. Um dos pontos-chave nesse processo é a redução de apresentação de conteúdo irrelevante, na busca, navegação, recomendação e interações entre usuários. A adoção de ferramentas capazes de auxiliar na filtragem de conteúdo indesejável, certamente trará grande benefício às comunidades virtuais.

O trabalho apresentado é o resultado da pesquisa desenvolvida para elaborar abordagens mais adequadas para a tarefa de classificação de vídeos em categorias semânticas. No escopo definido para este trabalho, essas categorias são vídeos não colaborativos do tipo *stuffing* e *spam*. Com os resultados gerados poderão ser desenvolvidas outras ferramentas que venham a auxiliar na busca e seleção de vídeos na *web* com base no conteúdo visual, incluindo ferramenta de anotação automática e correção de metadados.

Os experimentos realizados demonstram a boa acurácia da solução modelo, desenvolvida segundo técnicas mais tradicionais, no que diz respeito a identificação de elementos do tipo *stuffing*. Contudo, o difícil desafio do fosso semântico, aliado à forte

dependência do contexto, se apresentou como sendo uma barreira difícil de ser superada apenas com o uso dessas técnicas, para o problema do *spam*.

Dessa forma, desenvolveu-se uma representação inovadora, sensível ao contexto, elaborada para atender ao problema da identificação de vídeos do tipo *spam*, com resultados encorajadores. Uma vantagem dessa representação é a capacidade de generalização do conhecimento acerca do comportamento de vídeos das classes legítimo e *spam*, sugerida pelos resultados dos experimentos com a base controlada. Com isso, é possível a construção de ferramentas de detecção de vídeos do tipo *spam*, mesmo de posse de conjuntos de treinamento relativamente pequenos.

8.2 Trabalhos Futuros

As taxas de classificação encontradas neste trabalho ainda não autorizam uma abordagem de classificação totalmente automática. Uma alternativa possível é a sua utilização como uma etapa de filtragem de elementos suspeitos para posterior avaliação por especialistas humanos. É importante ressaltar que, propositalmente, neste trabalho foi realizado um esforço considerável para se utilizar apenas a informação visual, no intuito de demonstrar a importância desta na realização da tarefa de identificação de vídeos não colaborativos. Contudo, a partir da elaboração deste trabalho são apresentadas algumas sugestões ou desdobramentos para os resultados ora observados, que visam ampliar a qualidade das taxas de classificação, além de possibilitar o uso do conceito aqui construído em outras áreas de aplicação:

- **Análise multimodal:** conforme indicado no experimento de validação da Hipótese 1.3, Seção 5.7, e nos trabalhos elaborados por Benevenuto et al. [2009] e Langbehn et al. [2010] existem outras fontes de informação, que não apenas as puramente visuais, capazes de endereçar soluções ao problema de detecção de vídeos não colaborativos. Uma exploração combinada dessas fontes de informação pode contribuir consideravelmente para aumentar a acurácia dessa tarefa. São exemplos de fontes de informação: visual (imagem e movimento), áudio, texto (*captions*, comentários e metadados), relações e demais características sociais dos usuários que postam e acessam o vídeo;

- **Aumento do número de pontos de referência na *thread*:** a abordagem apresentada neste trabalho prevê a existência de apenas um ponto-de-referência na *thread* a ser utilizado para avaliar a adequação ou não do vídeo ao contexto da *thread*. O vídeo-original tem sido a escolha natural para essa aplicação. Contudo, a alternativa de identificação de mais pontos-de-referência em uma *thread* pode ajudar a resolver situações tais como a de um vídeo legítimo que possui conteúdo visual diverso ao do vídeo original, contudo abordam um mesmo conceito semântico. Nessa situação esse vídeo legítimo tende a apresentar grandes valores no vetor de distâncias, o que leva o classificador a confundi-lo com um vídeo do tipo *spam*. Para tal, faz-se necessário uma investigação detalhada do comportamento intrathread dos vídeos no intuito de observar um parâmetro que determine o instante em que deve ser adicionado mais pontos-de-referência em uma *thread* e quais os critérios para seleção deste novo ponto;
- **Exploração de espaço paramétrico:** como citado ao longo desse relatório, não se explorou variações de parâmetros nas técnicas acessórias à construção da representação sensível ao contexto. Obviamente, foram realizadas algumas observações no sentido de nortear o rumo a ser seguido, contudo essas não foram exauridas. Com isso, a realização desse tipo de investigação, identificando possíveis melhores combinações de parâmetros para configuração dessas técnicas também pode contribuir com a melhoria da qualidade da classificação. Entre outros pontos a serem considerados podemos citar: a utilização de técnica de resumo estático de vídeo, visto que a literatura dessa área tem avançado; a abordagem para construção do vocabulário visual, tanto no que diz respeito ao número de palavras visuais quanto ao uso da seleção aleatória; a abordagem para seleção de tópicos latentes relevantes, visto que o uso de todos os tópicos com valor singular maior que zero não é necessariamente a única alternativa para essa tarefa; a investigação de uma métrica mais adequada que a distância euclidiana para ser utilizada como métrica de distância no espaço de tópicos; a avaliação de outros *kernels* para o classificador SVM, bem como a exploração dos parâmetros de configuração desses; ainda, podem ser avaliadas outras possibilidades para a construção da base de dados *controlada*, observando o comportamento da

metodologia na classificação de *threads* desbalanceadas em relação a quantidade de elementos legítimo e *spam*;

- **Avaliação do SVD-parcial:** uma das dificuldades em se adotar a metodologia proposta em ambientes reais certamente é o custo envolvido para o cálculo/recálculo do SVD a cada novo elemento a ser classificado. Essa imposição se dá devido ao modo padrão de cálculo do SVD, que necessita conhecer todo o conjunto de dados (treinamento e testes) para obtenção dos valores singulares associados. A avaliação do impacto do custo/benefício da adoção de abordagens alternativas como o uso do SVD-parcial, conforme descrito por Tougas e Spiteri [2008], é uma verificação que se fará fundamental quando da decisão pela adoção da metodologia proposta em ambientes reais;
- **Aplicação em imagens e texto:** o conceito desenvolvido neste trabalho, que se remete a encontrar dentro do próprio contexto onde o vídeo está inserido as referências para determinar se este é ou não colaborativo, pode ser extrapolado para outros domínios, tais como: imagens e texto. A aplicação em comunidades virtuais que se destinam ao compartilhamento de imagens (por exemplo, Flickr e Pinterest) talvez seja de visualização mais direta, pois lida com elementos que podem ser representados por características visuais, tal qual neste trabalho. Contudo, abstraindo o tipo de característica utilizada para representar os elementos e entendendo que estão dispostos em conjuntos dentro da comunidade, o que dá uma noção do seu contexto semântico, o conceito desenvolvido neste trabalho também pode ser utilizado para analisar a ocorrência de mensagens textuais (tal como comentários em *blogs*, comunidades virtuais, entre outros) não colaborativas com o ambiente onde foram postadas.

8.3 Trabalhos Publicados

Até o presente momento foram publicados os seguintes artigos:

- A. da Luz, E. Valle e A. de A. Araujo. “***Semantic concepts for content filtering on video sharing social networks***”. Proceedings of the IADIS International

Conference WWW/Internet, ICWI, IADIS Press, Rio de Janeiro, RJ, Brazil, pp 293-300, 2011. (Extrato Qualis: **B2**);

- A. da Luz, E. Valle e A. de A. Araujo. “***Non-Collaborative Content Detecting on Video Sharing Social Networks***”. Journal of Multimedia Tools and Applications, Springer, 2012. (Extrato Qualis: **B2**);
- A. da Luz, E. Valle e A. de A. Araujo. “***A context-aware description for content filtering on video sharing social networks***”. IEEE International Conference on Multimedia and Expo, Melbourne, Australia, Julho, 2012. (Extrato Qualis: **A1**).

Referências

[Almeida et al., 2012] J. Almeida, N. J. Leite e R. da S. Torres. VISON: Video Summarization for ONline applications. *Pattern Recognition Letters*, v. 33, n. 4, SI, p. 397-409, MAR 2012.

[Androutsopoulos et al., 2000] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C. D. Spyropoulos, e P. Stamatopoulos. Learning to filter spam e-mail: A comparison of a Naïve Bayesian and a memory-based approach. *Proceedings of the workshop: Machine Learning and Textual Information Access*, pp. 1-13, 2000.

[Arbib, 1995] Arbib, Michael A. (Eds.). *The Handbook of Brain Theory and Neural Networks*, 1995.

[Avila et al., 2008a] S.E.F. Avila, A. Luz Jr. e A. de A. Araújo. Vsumm: A simple and efficient approach for automatic video summarization. *International Conference on Systems, Signals and Image Processing*, 2008.

[Avila et al., 2008b] S.E.F. Avila, A. Luz Jr., A. de A. Araújo e M. Cord. VSUMM: An approach for automatic video summarization and quantitative evaluation. *Brazilian Symposium on Computer Graphics and Image Processing - SIBGRAPI*, Campo Grande, MS, Brazil, pp 103-110, 2008.

[Backstrom et al., 2006] L. Backstrom, D. Huttenlocher, J. Kleinberg e X. Lan. Group formation in large social networks: membership, growth, and evolution. *ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 44--54, Philadelphia, PA, USA, 2006.

[Ballan et al., 2010] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari e G. Serra. Event detection and recognition for semantic annotation of video. *Journal of Multimedia Tools and Applications*, Springer, vol. 51, pp. 279-302, 2010.

[Batista et al., 2009] N. Batista, A. Lopes e A. Araújo. Detecting Buildings in Historical Photographs Using Bag-of-Keypoints. *Brazilian Symposium on Computer Graphics and Image Processing - SIBGRAPI*, 2009.

[Belkin e Niyogi, 2001] M. Belkin e P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Journal of Neural Information Processing Systems*, 2001.

[Benevenuto et al., 2009] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida e M. Gonçalves. Detecting spammers and content promoters in online video social networks. *ACM SIGIR conference on Research and development in information retrieval*, pp. 620--627, Boston, MA, Julho, 2009.

[Benevenuto et al., 2012] F. Benevenuto, T. Rodrigues, A. Veloso, J. Almeida, M. Gonçalves e V. Almeida. Practical Detection of Spammers and Content Promoters in Online Video Sharing Systems. *IEEE Transactions on Systems, Man and Cybernetics - Part B*, **to appear**, 2012.

[Benevenuto, 2009] F. Benevenuto. Pollution in online video social networks. Universidade Federal de Minas Gerais (*tese de doutorado*), 2009.

[Benmokhtar e Huet, 2006] R. Benmokhtar e B. Huet. Classifier Fusion: Combination Methods For Semantic Indexing in Video Content. *Int. Conference on Artificial Neural Networks* [2]: 65-74, 2006.

[Biggio et al., 2011] B. Biggio, G. Fumera, I. Pillai e F. Roli. A survey and experimental evaluation of image spam filtering techniques. *Pattern Recogn. Lett.* 32, 10, 1436-1446, 2011.

[Bíró e Szabó, 2009] I. Bíró e J. Szabó. Latent Dirichlet Allocation for Automatic Document Categorization. *European Conference on Machine Learning and 12th Principles of Knowledge Discovery in Databases*, 2009.

[Blei et al., 2003] D. Blei, A. Ng, M. Jordan, J. Lafferty. Latent Dirichlet allocation. *Journal of Machine Learning Research*: pp. 993–1022, 2003.

[Boureau et al., 2010] Y-L. Boureau, F. Bach, LeCun Y. e J. Ponce. Learning Mid-Level Features For Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[Brezeale e Cook, 2008] D. Brezeale e D. J. Cook. Automatic Video Classification: A

Survey of the Literature. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol.38, no.3, pp.416-430, May, 2008.

[Burges, 1998] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov., vol. 2, no. 2, pp. 121–167, 1998.

[Burgess e Green, 2009] J. Burgess e J. Green. YouTube: Online Video and Participatory Culture. Polity, 2009.

[Caruana e Niculescu-Mizil, 2006] R. Caruana e A. Niculescu-Mizil. An empirical comparison of supervised learning algorithms. Proceedings of the 23rd international conference on Machine learning, 2006.

[Castillo et al., 2007] C. Castillo, D. Donato, A. Gionis, V. Murdock e F. Silvestri. Know your neighbors: web spam detection using the web topology. ACM SIGIR Conference on Research and development in Information Retrieval, pp. 423--430, New York, NY, USA. ACM, 2007.

[Cha et al., 2007] M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahn e S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. ACM SIGCOMM Conference on Internet Measurement, pp. 1--14, New York, NY, USA, 2007.

[Chang e Chen, 2007] I.-C. Chang e K.-Y. Chen. Content-selection based video summarization. International Conference on Consumer Electronics, pp. 1-2, 2007.

[Chang e Lin, 2011] C.-C. Chang e C.-J. Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, vol. 2, 3, 1--27, 2011.

[Chang et al., 1998] E. Y. Chang, J. Z. Wang, C. Li e G. Wiederhold. RIME: A Replicated Image Detector for the World-Wide Web. Multimedia storage and archiving systems III, Boston, MA, USA, November, 1998.

[Colombino et al., 2010] T. Colombino, D. Martin, A. Grasso e L. Marchesotti. A Reformulation of the Semantic Gap Problem in Content-Based Image Retrieval Scenarios. Int. Conf. on the Design of Cooperative Systems. France - 19-21 May, 2010.

[Coppin et al., 2007] G. Coppin, F. Cadier e P. Lenca. Some considerations on cognitive modeling for collective decision support. Hawaii International Conference on System

Sciences - HICSS, Hawaii, 2007.

[Cotsaces et al., 2006] C. Cotsaces, N. Nikolaidis e L. Pitas. Video shot detection and condensed representation: A review. *IEEE Signal Processing Magazine*, 23(2):28—37, 2006.

[Crane e Sornette, 2008] R. Crane e D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system, *Proc. Nat. Acad. Sci. USA* 105 (41), 15649-15653, 2008

[Crane et al., 2008] R. Crane, F. Schweitzer e D. Sornette. Power law signature of media exposure in human response waiting time distributions. *Physical Review E*, 1(5):056101+, May 2010.

[Cunha et al., 2012] T. O. Cunha, F. Souza, A. de A. Araújo e G. L. Pappa. Rushes Video Summarization Based on Spatio-temporal Features. *ACM Symposium On Applied Computing*, Riva del Garda (Trento), Italy, 2012.

[Cybenko, 1989] G. Cybenko. Approximations by superposition of a sigmoidal function. *Mathematics of Control, Signal and Systems*, vol. 2, pp. 303—314, 1989.

[Datta et al., 2008] R. Datta, D. Joshi, J. Li e J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.* 40, 2, Apr. 2008.

[Deselaers et al., 2008] T. Deselaers, L. Pimenidis e H. Ney. Bag-of-visual-words models for adult image classification and filtering. *International Conference on Pattern Recognition*, pp. 1-4, 2008.

[Ding et al., 2011] Y. Ding, J. Zhang, J. Li e X. Wei. A Bag-of-Feature Model for Video Semantic Annotation. *International Conference on Image and Graphics*, pp. 696-701, 2011.

[Drucker et al., 1999] H. Drucker, D. Wu e V.N. Vapnik. Support vector machines for spam categorization. *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 1048-1054, 1999.

[Fa, 1967] Fa, H. Handbook of the poisson distribution, 1967.

[Fagni et al., 2010] T. Fagni, F. Falchi e F. Sebastiani. Image classification via adaptive ensembles of descriptor-specific classifiers. *Pattern Recognition and Image Analysis*, vol. 20, no. 1, pp. 21-28, March, 2010.

[Fergus et al., 2003] R. Fergus, P. Perona e A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 264–271, 2003.

[Fergus et al., 2010] R. Fergus, L. Fei-Fei, P. Perona e A. Zisserman. Learning Object Categories From Internet Image Searches. *Proceedings of the IEEE*, vol.98, no.8, pp.1453-1466, Aug. 2010.

[Ferragina, 2011] P. Ferragina. Beyond the bag-of-words paradigm to enhance information retrieval applications. *International Conference on Similarity Search and Applications - SISAP*. ACM, New York, NY, USA, 3-4, 2011.

[Fetterly et al., 2004] D. Fetterly, M. Manasse e M. Najork. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. *International Workshop on the Web and Databases*, pp. 1--6, New York, NY, USA. ACM, 2004.

[Freud e Schapire, 1996] Y. Freud e R. Schapire. Experiments with a new boosting algorithms. *International Conference on Machine Learning*, 1996.

[Furini et al., 2007] M. Furini, F. Geraci, M. Montangelo e M. Pellegrini. VISTO: visual storyboard for web video browsing. *ACM International Conference on Image and Video Retrieval (CIVR)*, pp. 635—642, 2007.

[Gerard e Buckley, 1988] S. Gerard e C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24 (5): 513–523, 1988.

[Gomes et al., 2007] L. H. Gomes, C. Cazita, J. M. Almeida, V. Almeida e W. M. Meira, Jr.. Workload models of spam and legitimate e-mails. *Perform. Eval.*, 64(7-8):690—714, 2007.

[Gonzalez e Woods, 2006] R. C. Gonzalez e R. E. Woods. *Digital Image Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 3 edição, 2006.

[Goyal e Walia, 2012] A. Goyal e E. Walia. An Analysis of Shape Based Image Retrieval Using Variants of Zernike Moments as Features. *International Journal of Imaging*, vol. 7, 1, 2012.

[Guillaumin et al., 2010] M. Guillaumin, J. Verbeek, C. Schmid. Multimodal semi-

supervised learning for image classification. IEEE Conference on Computer Vision & Pattern Recognition, Jun, 2010.

[Gyöngyi et al., 2004] Z. Gyöngyi, H. Garcia-Molina e J. Pedersen. Combating web spam with trustrank. International Conference on Very large Data Bases - VLDB, pp. 576--587, Toronto, Canada, 2004.

[Hadi et al., 2006] Y. Hadi, F. Essannouni e R. O. H. Thami. Video summarization by k-medoid clustering. ACM Symposium on Applied Computing - SAC, pp. 1400--1401, New York, NY, USA, 2006.

[Han, 2009] D. Han, L. F. Bo e C. Sminchisescu. Selection and context for action recognition. IEEE International Conference on Computer Vision. Kyoto, Japan, 2009.

[Heymann et al., 2007] P. Heymann, G. Koutrika e H. Garcia-Molina. Fighting spam on social web sites: A survey of approaches and future challenges. IEEE Internet Computing, 11[6]:36—45, 2007.

[Hofmann, 1999] T. Hofmann. Probabilistic latent semantic indexing. International ACM SIGIR Conference on Research and Development in information Retrieval. Berkeley, California, August, 1999.

[Hosmer e Lemeshow, 2000] D. W. Hosmer e S. Lemeshow. Applied Logistic Regression, 2nd ed.. New York; Chichester, Wiley, 2000.

[Hu et al., 2010] F. Hu, L. Luo, F. Zhang e J. Liu. Action recognition using hybrid spatio-temporal bag-of-features. International Conference on Computer Sciences and Convergence Information Technology - ICCIT, vol., no., pp.812-815, 2010.

[Irani et al., 2010] D. Irani, S. Webb e C. Pu. Study of Static Classification of Social Spam Profiles in MySpace. Int. AAAI Conf. on Weblogs and Social Media, May 23-26, 2010.

[Jégou et al., 2010] H. Jégou, M. Douze e C. Schmid. Improving Bag-of-Features for Large Scale Image Search. International Journal of Computer Vision, Springer, vol. 87, 3, pp. 316-336, 2010.

[Jiang et al., 2007] Y.-G. Jiang, C.-W. Ngo e J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. ACM International Conference on

Image and Video Retrieval, pp. 494—501, Amsterdam, The Netherlands, 2007.

[Jiang, 2010] W. Jiang. Advanced Techniques for Semantic Concept Detection in General Videos. Tese de Doutorado em Artes e Ciências na Universidade de Columbia, 2010.

[Jindal e Liu, 2008] N. Jindal e B. Liu. Opinion spam and analysis. International Conference on Web Search and Web Data Mining, pp. 219—230, New York, NY, USA, 2008.

[Joshi et al., 2005] M. Joshi, T. Pedersen e R. Maclin. A comparative study of support vector machines applied to the supervised word sense disambiguation problem in the medical domain. Indian International Conference on Artificial Intelligence, 2005.

[Jurie e Triggs, 2005] F. Jurie e B. Triggs. Creating efficient codebooks for visual recognition. IEEE International Conference on Computer Vision, Beijing, China, pp. 604—610, 2005.

[Kato, 1992] T. Kato. Database architecture for content-based image retrieval. Proceedings of SPIE Image Storage and Retrieval Systems, 1662:112-123, 1992.

[Ke e Sukthankar, 2004] Y. Ke e R. Sukthankar. Pca-sift: a more distinctive representation for local image descriptors. IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pp. 506—513, 2004.

[Ke et al., 2004] Y. Ke, R. Sukthankar e L. Huston. An efficient parts-based near-duplicate and sub-image retrieval system. ACM International Conference on Multimedia, New York, NY, USA, October, 2004.

[Kherfi et al., 2004] M. L. Kherfi, D. Ziou e A. Bernardi. Image Retrieval from the World Wide Web: Issues, Techniques, and Systems. ACM Comput. Surv. 36, 1, Mar., 2004.

[Khoshabeh e Hollan, 2009] R. Khoshabeh e J. D. Hollan. Spatio-temporal interest points for video analysis. International Conference Extended Abstracts on Human Factors in Computing Systems, Boston, MA, USA, April, 2009.

[Khotanzad e Hong, 1990] A. Khotanzad e Y. H. Hong. Invariant image recognition by zernike moments. IEEE Trans. Pattern Anal. Mach. Intell., 12(5):489—497, 1990.

[Kimura et al., 2011] P. A. S. Kimura, J. M. B. Cavalcanti, P. C. Saraiva, R. S. Torres, M. A. Gonçalves. Evaluating Retrieval Effectiveness of Descriptors for Searching in Large Image Databases. *Journal of Information and Data Management*, vol 2, 3, 2011.

[Kinnunen et al., 2009] T. Kinnunen, J.-K. Kamarainen, L. Lensu e H. Kälviäinen. Bag-of-features codebook generation by self-organisation. *International Workshop on Advances in Self-Organizing Maps*, pp. 124--132, Berlin, Heidelberg. Springer-Verlag, 2009.

[Koprinska et al., 2007] I. Koprinska, J. Poon, J. Clark e J. Chan. Learning to classify e-mail. *Inf. Sci.* 177 (10), 2167–2187, 2007.

[Kuncheva et. al., 2001] L. Kuncheva, J.C. Bezdek e R. Duin. Decision templates for multiple classifier fusion: an experiemental comparaison. *Pattern Recognition*, vol. 34, pp. 299–314, 2001.

[Lai et al., 2009] C.-C. Lai, C.-H. Wu e M.-C Tsai. Feature selection using particle swarm optimization with application in spam filtering. *Int. Journal of Innovative Comput.* 5 (2), 423–432, 2009.

[Lajevardi e Hussain, 2009] S. M. Lajevardi e Z. M. Hussain. Zernike Moments for Facial Expression Recognition. *IEEE Int. Conf. On Communication Computer and Power*, pp. 378-381, Omã, Fev., 2009.

[Landauer et al., 1998] T. Landauer, P. Foltz e D. Laham. Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259-284, 1998.

[Langbehn et al., 2010] H. Langbehn, S. Ricci, M. Gonçalves, J. Almeida, G. Pappa e F. Benevenuto. A Multi-view Approach for Detecting Non-Cooperative Users in Online Video Sharing Systems. *J. of Inf. and Data Management*, v.1, 2010.

[Laptev et al., 2008] I. Laptev, M. Marszalek, C. Schmid e B. Rozenfeld. Learning realistic human actions from movies. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, pp. 1-8, 2008.

[Laptev, 2005] I. Laptev. On Space-Time Interest Points. *Int. J. Comput. Vision* 64, 2-3 (Sep. 2005), 107-123, 2005.

[Le et al., 2007] C. A. Le, V.-N. Huynh, A. Shimazu e Y. Nakamori. Combining classifiers

for word sense disambiguation based on dempster-shafer theory and owa operators. *Data & Knowledge Engineering*, vol. 63, no. 2, pp. 381-396, November, 2007.

[Lee et al., 2010] K. Lee, J. Caverlee e S. Webb. The Social Honeypot Project: Protecting Online Communities From Spammers. *Int. World Wide Web Conference*, April 26-30, 2010.

[Leung e Malik, 2001] T. K. Leung e J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, vol. 43, no. 1, pp. 29–44, 2001.

[Li et al., 2001] Y. Li, T. Zhang e D. Tretter. An overview of video abstraction techniques. Technical report, HP Laboratory, HP-2001-191, 2001.

[Li et al., 2009] L. Li, L. Duan, Q. Huang, J. Du e W. Gao. A generic approach to classify sports video shots and its application in event detection. *International Conference on Internet Multimedia Computing and Service*. Kunming, Yunnan, China, November, 2009.

[Lin e Hauptmann, 2002] W. Lin e A. Hauptmann. News video classification using SVM-based multimodal classifiers and combination strategies. *ACM International Conference on Multimedia*, Juan-les-Pins, France, December 01 - 06, 2002.

[Lin et al., 2003] Y.-C. Lin, H.-W. Tseng e C.-S. Fuh. Pornography Detection Using Support Vector Machine. *Conference on Computer Vision, Graphics and Image Processing - CVGIP*, pp. 123-130, Kinmen, Taiwan, 2003.

[Lin et al., 2008] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura e B. L. Tseng. Detecting splogs via temporal dynamics using self-similarity analysis. *ACM Trans. Web*, 2(1):1—35, 2008.

[Liu et al., 2006] S. Liu, M. Xu, H. Yi, L. Chia e D. Rajan. Multimodal semantic analysis and annotation for basketball video. *EURASIP Journal Appl. Signal Process*. Jan., 2006.

[Liu et al., 2007] Y. Liu, D. Zhang, G. Lu e W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262—282, 2007.

[Liu et al., 2008] J. Liu, S. Ali e M. Shah. Recognizing human actions using multiple features. *IEEE Conference on Computer Vision and Pattern Recognition – CVPR*, pp. 1-8,

2008.

[Lopes et al., 2009a] A. P. B. Lopes, S. E. F. Avila, A. N. A. Peixoto, R. S. Oliveira e A. de A. Araújo. A bag-of-features approach based on Hue-SIFT descriptor for nude detection. European Signal Processing Conference – EUSIPCO, Glasgow, Scotland, pp. 1552-1556, 2009.

[Lopes et al., 2009b] A. P. B. Lopes, R. S. Oliveira, J. M. Almeida e A. de A. Araújo. Comparing alternatives for capturing dynamic information in bag-of-visual-features approaches applied to human actions recognition. IEEE International Workshop on Multimedia Signal Processing – MMSP, Rio de Janeiro, RJ, Brazil, 6p, 2009.

[Lopes et al., 2009c] A. P. B. Lopes, R. S. Oliveira, J. M. Almeida e A. de A. Araújo. Spatio-temporal frames in a bag-of-visual-features approach for human actions recognition. Brazilian Symposium on Computer Graphics and Image Processing – SIBGRAPI, Rio de Janeiro, RJ, Brazil, pp 315-321, 2009.

[Lopes et al., 2011] A. P. B. Lopes, E. Santos, E. A. Valle Jr., J. M. Almeida e A. de A. Araújo. Transfer learning for human action recognition. Proceedings of the 24th Brazilian Conference on Graphics, Patterns and Images – SIBGRAPI, Maceió, AL, Brazil, pp 352-359, 2011.

[Lowe, 1999] D. G. Lowe. Object recognition from local scale-invariant features. IEEE International Conference on Computer Vision, vol. 2, p. 1150, Washington, DC, USA, 1999.

[Luz et al., 2011] A. da Luz, E. Valle e A. de A. Araujo. Semantic concepts for content filtering on video sharing social networks. IADIS International Conference WWW/Internet, ICWI, Rio de Janeiro, RJ, Brazil, pp 293-300, 2011.

[Ma e Zhang, 1998] W.-Y. Ma, H. Zhang. Structured and content-based video browsing. Conference Record of Thirty Second Asilomar Conference on Signals Systems and Computers, 1, 253-257, 1998.

[Mason et al., 2000] L. Mason, J. Baxter, P. Bartlett e M. Frean. Boosting algorithms as gradient descent. In S.A. Solla, T.K. Leen, and K.-R. Muller, editors, Advances in Neural Information Processing Systems 12, pages 512--518. MIT Press, 2000.

[Matsui e Tanabe, 2006] T. Matsui e K. Tanabe. Comparative Study of Speaker Identification Methods: dPLRM, SVM and GMM. IEICE - Trans. Inf. Syst. E89-D, 3, Mar. 2006.

[Mehta et al., 2008] B. Mehta, S. Nangia, M. Gupta e W. Nejdl. Detecting image spam using visual features and near duplicate detection. Int. Conf. on World Wide Web. ACM, New York, NY, USA, 497-506, 2008.

[Mikolajczyk e Schmid, 2005] K. Mikolajczyk e C. Schmid. A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell., 27(10):1615—1630, 2005.

[Mikolajczyk et al., 2005] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir e L. Van Gool. A Comparison of Affine Region Detectors. Int. J. Comput. Vision 65, 1-2, 43-72, 2005.

[Monay e Gatica-Perez, 2007] F. Monay e D. Gatica-Perez. Modeling Semantic Aspects for Cross-Media Image Indexing. IEEE Trans. Pattern Anal. Mach. Intell. 29, 10, 1802-1817, Oct., 2007.

[Mundur et al., 2006] P. Mundur, Y. Rao e Y. Yesha. Keyframe-based video summarization using Delaunay clustering. International Journal on Digital Libraries, 6[2]:219—232, 2006.

[Naphade e Yeo, 2000] M. Y. M. Naphade e B.-L. Yeo. A novel scheme for fast and efficient video sequence matching using compact signatures. Storage and Retrieval for Media Databases – SPIE, Vol. 3972, pp. 564–572, Jan. 2000.

[Newman e Park, 2003] M. E. J. Newman e J. Park. Why social networks are different from other types of networks. Journal of Phys. Rev., vol. 68, 3, 2003.

[Ng e Jordan, 2001] A. Ng e M. Jordan. On Discriminative vs. Generative classifiers: A comparison of Logistic Regression and Naïve Bayes. Neural Information Processing Systems, 2001.

[Niebles et al., 2006] J. C. Niebles, H. Wang e L. Fei-Fei. Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. Int. J. Comput. Vision 79, 3, Sep. 2006.

[Paalanen et al. 1995] P. Paalanen, J. Kamarainen, J. Ilonen e H. Kalviainen. Feature representation and discrimination based on gaussian mixture model probability densities. Research Report, Lappeenranta University of Technology, 1995.

[Paleari et al., 2009] M. Paleari, R. Benmokhtar e B. Huet. Evidence Theory-Based Multimodal Emotion Recognition. International Multimedia Modeling Conference on Advances in Multimedia Modeling. Lecture Notes In Computer Science, vol. 5371. Springer-Verlag, Berlin, Heidelberg, 435-446, 2009.

[Pearson, 1901] K. Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine (6): 559–572, 1901.

[Penatti et al., 2012] O. A. B. Penatti, E. Valle e R. S. Torres. Comparative study of global color and texture descriptors for web image retrieval. J. Vis. Comun. Image Represent. 23, 2, 359-380, 2012.

[Permuter et al., 2003] H. Permuter, J. Francos e I. H. Jermyn. Gaussian mixture models of texture and colour for image database retrieval. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.

[Pless e Souvenir, 2009] R. Pless e R. Souvenir. A Survey of Manifold Learning for Images. IPSJ Transactions on Computer Vision and Applications, Vol. 1, pp.83-94, 2009.

[Polikar, 2006] R. Polikar. Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, 6(3):21–45, 2006.

[Poppe, 2010] R. Poppe. A survey on vision-based human action recognition. Journal of Image Vision Computing. 28, 6, 976-990, 2010.

[Rabiner, 1989] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE vol. 77(2): 257–286, Feb., 1989.

[Ronda et. al., 2008] T. F. Ronda, P. M. A. Marques, J. C. Felipe. Avaliação Comparativa de descritores de forma utilizados para caracterização e classificação de nódulos de mama. 21º Congresso Brasileiro de Engenharia Biomédica, Salvador-BA, Novembro, 2008.

[Roweis e Lawrance, 2000] S. T. Roweis e K. S. Lawrance. Nonlinear Dimensionality

reduction by locally linear embedding. *Science*, 290, pp. 2323-2326.

[Ruta e Gabrys, 2000] D. Ruta e B. Gabrys. An overview of classifier fusion methods. *Computing and Information Systems Journal* 7(1), 1-10, 2000.

[Ruta e Gabrys, 2001] D. Ruta e B. Gabrys. Analysis of the correlation between majority voting error and the diversity measures in multiple classifier systems. *International Symposium on Soft Computing*, Paisley, UK, 2001.

[Ruta e Gabrys, 2002] D. Ruta e B. Gabrys. A theoretical analysis of the, limits of majority voting errors for multiple classifier systems. *Pattern Analysis and Applications* 5 [4], 2002.

[Ruta e Gabrys, 2005] D. Ruta e B. Gabrys. Classifier Selection for Majority Voting. *Information Fusion*, [6], 63-81, 2005.

[Sannen et al., 2010] D. Sannen, E. Lughofer e H. Van Brussel. Towards incremental classifier fusion. *Intell. Data Anal.* 14, 1, Jan, 2010.

[Schmid e Mohr, 1996] C. Schmid e R. Mohr. Combining gray-value invariants with local constraints for object recognition. *Conference on Computer Vision and Pattern Recognition*, pp. 872–877, 1996.

[Shabani et al., 2009] H. Shabani, D. A. Clausi e J. S. Zelek. Towards a Robust Spatio-Temporal Interest Point Detection for Human Action Recognition. *Canadian Conference on Computer and Robot Vision*, May, 2009.

[Sharkey e Sharkey, 1997] A. J. C. Sharkey e N. E. Sharkey. Combining diverse neural nets. *The Knowledge Engineering Review* 12 (3), 1997.

[Silovsky et al., 2009] J. Silovsky, P. Cerva e J. Zdansky. Comparison of Generative and Discriminative Approaches for Speaker Recognition with Limited Data. *Radio Engineering*, vol. 18, n. 3, Set., 2009.

[Singh et al., 2011] C. Singh, E. Walia e N. Mittal. Rotation invariant complex Zernike moments features and their applications to human face and character recognition. *Computer Vision, IET* , vol.5, no.5, pp.255

[Sivic e Zisserman, 2003] J. Sivic e A. Zisserman. Video Google: A text retrieval

approach to object matching in videos. IEEE International Conference on Computer Vision, Nice, France, 2003, pp. 1470–1477.

[Sivic e Zisserman, 2008] J. Sivic e A. Zisserman. Efficient visual search for objects in videos. *Proceedings of the IEEE*, 96(4):548–566, 2008.

[Skurichina e Duin, 1998] M. Skurichina e R. Duin. Bagging for linear classifiers. *Pattern Recognition*, vol. 31, no. 7, pp. 909–930, 1998.

[Smeulders et. al., 2000] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta e R. Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans Pattern Anal Mach Intell*, 22(12):1349-80, 2000.

[Snoek e Worring, 2009] C. G. M. Snoek e M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215-322, 2009.

[Snoek et. al., 2005] C. G. Snoek, M. Worring e A. W. Smeulders. Early versus late fusion in semantic video analysis. *ACM International Conference on Multimedia*. Hilton, Singapore, November 06 - 11, 2005.

[Song et al., 2011] Z. Song, Q. Chen, Z. Huang, Y. Hua e S. Yan. Contextualizing object detection and classification. *IEEE Conference on Computer Vision and Pattern Recognition [CVPR]*, vol., no., pp.1585, 2011.

[Souza et al., 2010] F. D. M. Souza, G. Cámara-Chávez, E. A. Valle Jr. e A. de A. Araújo. Violence detection in video using spatio-temporal features. *Brazilian Conference on Graphics, Patterns and Images – SIBGRAPI*, Gramado, RS, Brazil, pp 224-230, 2010.

[Souza et al., 2011] F. D. M. Souza, E. A. Valle Jr., G. Cámara-Chávez e A. de A. Araújo. Color-aware local spatiotemporal features for action recognition. *Iberoamerican Congress on Pattern Recognition – CIARP*, vol. 7042, Pucón, Chile, pp 248-255, 2011.

[Souza Jr. et al., 2011] S. F. Souza Jr., A. de A. Araújo e D. Menotti. An overview of automatic event detection in soccer matches. *IEEE Workshop on Applications of Computer Vision – WACV*, Kona, Hawaii, USA, pp 31-38, 2011.

[Spyrou et al., 2009] E. Spyrou, G. Toliás, P. Mylonas e Y. S. Avrithis. Concept detection and keyframe extraction using a visual thesaurus. *Multimedia Tools and Applications*,

41[3]: 337-373, 2009.

[Tenenbaum et al., 2000] J. Tenenbaum, V. de Silva e J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[Thomason, 2007] A. Thomason. Blog spam: A review. *Conference on e-mail and Anti-Spam (CEAS)*, 2007.

[Tian et al., 2011] Q. Tian, S. Zhang, W. Zhou, R. Ji, B. Ni e N. Sebe. Building descriptive and discriminative visual codebook for large-scale image applications. *Multimedia Tools and Applications*, vol. 51, 2, pp. 441-477, 2011.

[Torres e Falcão, 2006] R. da S. Torres e A. X. Falcão. Content-based Image Retrieval: Theory and Applications. *Revista de Informática Teórica e Aplicada*, 13, (2), 161–185, 2006.

[Tougas e Spiteri, 2008] J. E. Tougas e R. J. Spiteri. Two uses for updating the partial singular value decomposition in latent semantic indexing. *Appl. Numer. Math.* 58, 4, 499-510, 2008.

[Truong e Venkatesh, 2007] B. T. Truong e S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3(1), 2007.

[Tuytelaars e Mikolajczyk, 2008] T. Tuytelaars e K. Mikolajczyk. Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.* 3, 3 (July 2008), 177-280, 2008.

[Ulusoy e Bishop, 2005] I. Ulusoy e C. M. Bishop. Generative versus discriminative methods for object recognition. *IEEE Computer Vision and Pattern Recognition*, vol.2, no., pp. 258, 2005.

[Valle et al., 2008] E. Valle, M. Cord e S. Philipp-Foliguet. Fast identification of visual documents using local descriptors. *ACM Symposium on Document Engineering*, Sao Paulo, Brazil, September, 2008.

[Van de Sande, 2010] K. E. A. van de Sande, T. Gevers e C. G. M. Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[Van de Weijer et al., 2006] J. van de Weijer, T. Gevers e A. Bagdanov. Boosting color saliency in image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 150–156, 2006.

[Vapnik, 1979] V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Nauka, Moscow. [*in Russian*] (English translation: 1982, Springer Verlag, New York), 1979.

[Wall et al., 2003] M. Wall, A. Rechtsteiner, L. M. Rocha. “Singular value decomposition and principal component analysis”. In *A Practical Approach to Microarray Data Analysis*. D.P. Berrar, W. Dubitzky, M. Granzow, eds. pp. 91-109, Kluwer: Norwell, MA. LANL LA-UR-02-4001, 2003.

[Wang et al., 2009] H. Wang, M. Ullah, A. Klaser, I. Laptev e C. Schmid. Evaluation of local spatio-temporal features for action recognition. *British Machine Vision Conference*, September, 2009.

[Wang, 2010] A. H. Wang. Detecting spam bots in online social networking sites: a machine learning approach. *Working Conf. on Data and Applications Security and Privacy*, Springer-Verlag, Berlin, Heidelberg, 335-342, 2010.

[Wu e Tsai, 2009] C. Wu e C. Tsai. Robust classification for spam filtering by back-propagation neural networks using behavior-based features. *Appl. Int.* 31 (2), 107–121, 2009.

[Wu et al., 2005] C.-T. Wu, K.-T. Cheng, Q. Zhu e Y.-L. Wu. Using visual features for anti-spam filtering. *IEEE International Conference on Image Processing – ICIP*, vol. 3, pp. III 509 12, 2005.

[Wu et al., 2007] X. Wu, W. Zhao e C. Ngo. Near-duplicate keyframe retrieval with visual keywords and semantic context. *ACM International Conference on Image and Video Retrieval*. Amsterdam, The Netherlands, July, 2007.

[Wu et al., 2011] X. Wu, C.-W. Ngo, Y.-M. Zhu e Q. Peng. Boosting Web Video Categorization With Contextual Information from Social Web. *World Wide Web Journal*, vol. 15, 2, 197-212, 2011.

[Xue e Titterington, 2008] J. Xue e D. M. Titterington. On Discriminative vs. Generative

Classifiers: A Comparison of Logistic Regression and Naive Bayes. *Neural Process. Lett.* 28, 3, Dec., 2008.

[Yan e Naphade, 2005] R. Yan, M. Naphade. Multi-Modal Video Concept Extraction Using Co-Training. *IEEE International Conference on Multimedia and Expo*, July, 2005.

[Yang et al., 2007a] L. Yang, J. Liu, X. Yang e X. Hua. Multi-modality web video categorization. *International Workshop on Workshop on Multimedia information Retrieval*. Augsburg, Bavaria, Germany, September, 2007.

[Yang et al., 2007b] J. Yang, Y. Jiang, A. G. Hauptmann e C. Ngo. Evaluating bag-of-visual-words representations in scene classification. *ACM Multimedia Information Retrieval*. Augsburg, Bavaria, Germany, September 24 - 29, 2007.

[Yoshitaka e Ichikawa, 1999] A. Yoshitaka e T. Ichikawa. A survey on content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering*, 11[1]:81—93, 1999.

[You et al., 2010] J. You, G. Liu e A. Perkis. A semantic framework for video genre classification and event analysis. *Signal Processing: Image Communication*, Elsevier, 25, 4, 287-302, Apr., 2010.

[Yu et al., 2006] H. Yu, M. Kaminsky, P. B. Gibbons e A. Flaxman. Sybilguard: defending against sybil attacks via social networks. *Conference on Applications, technologies, Architectures, and Protocols for Computer Communications*, pp. 267--278, New York, NY, USA. ACM Press, 2006.

[Yuan et al., 2007] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin e B. Zhang. A formal study of shot boundary detection. *Circuits and Systems for Video Technology*, *IEEE Transactions on*, 17(2):168-186, 2007.

[Zernike, 1934] F. Zernike. Diffraction theory of the cut procedure and its improved form, the phase contrast method. *Physica*, 1:pp. 689-704, 1934.

[Zhang e Gong, 2010] J. Zhang e S. Gong. Action categorization by structural probabilistic latent semantic analysis. *Comput. Vis. Image Underst.* 114, 857-864.

[Zhang et al., 2007] J. Zhang, M. Marszalek, S. Lazebnik e C. Schmid. Local Features and

Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *Int. J. Comput. Vision* 73, 2, 213-238, 2007.

[Zhang et al., 2009] T. Zhang, H. Lu e S. Z. Li. Learning semantic scene models by object classification and trajectory clustering. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1940–1947, June, 2009.

[Zhang, 2004] H. Zhang. The optimality of naive bayes. In V. Barr and Z. Markov, editors, *FLAIRS Conference*. AAAI Press, 2004.

[Zhao e Pietikäinen, 2007] G. Zhao e M. Pietikäinen. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.29, no.6, pp.915-928, June 2007.

[Zhou et al., 2011] W. Zhou, Y. Lu, H. Li e Q. Tian. Latent visual context learning for web image applications. *Pattern Recognition*, Special issue: Semi-Supervised Learning for Visual Content Analysis and Understanding, 44(10-11): 2263-2273, 2011.

[Zhuang et al., 1998] Y. Zhuang, Y. Rui, T. S. Huang e S. Mehrotra. Adaptive key frame extraction using unsupervised clustering. *IEEE International Conference on Image Processing – ICIP*, vol. 1, pp. 866—870, 1998.

Glossário

Alguns dos termos utilizados ao longo deste trabalho são muito específicos ao domínio de aplicação. Por essa ser uma nova área de exploração, a definição e interpretação desses termos pode causar dúvidas na compreensão do texto. A seguir são definidos alguns termos conforme o entendimento adotado para este trabalho:

Bag-of-topic-differences (BoTD):	É uma abordagem de representação de informação, construída neste trabalho, que altera o espaço de projeção geométrica do <i>bag-of-visual-features</i> para o espaço de semântica latente, utilizando o SVD. Ver Seção 6.1;
Bag-of-topics (BoT):	É uma abordagem de representação de informação construída neste trabalho, resultante da mudança do espaço de projeção geométrica do <i>bag-of-visual-features</i> para o espaço de semântica latente, utilizando o SVD. Ver Seção 6.1;
Bag-of-visual-differences (BoVD):	É uma abordagem de representação de informação construída neste trabalho, baseada no <i>bag-of-visual-features</i> , com habilidade de considerar o contexto no qual o vídeo está inserido. Ver Seção 6.1;
Bag-of-visual-features (BoVF):	É uma abordagem tradicional de representação de informação, capaz de elevar a um médio-nível semântico um vetor de características construído com informações de baixo-nível. Ver Seção 3.7;
Base Controlada:	Base de dados sintética criada para possibilitar maior conhecimento e controle do modo como os vídeos estão distribuídos nas <i>threads</i> (contextos e variabilidade inter e intrathread). Ver Seção 7.1.1;
Base Selvagem:	Base de dados que representa o ambiente real de uma rede de social de compartilhamento de vídeo. Ver Seção 5.1.1;
Contexto de um vídeo:	Refere-se à semântica da informação visual apresentada pelos vídeos originais e legítimos dentro de uma thread. É entendido que um vídeo está relacionado ao contexto inserido, quando este possui temática similar aos demais vídeos original e legítimos na <i>thread</i> ;
Estatística de quadro:	É uma característica não visual elaborada neste trabalho e utilizada para representar um vídeo. É obtida com base no resultado do resumo estático gerado para o vídeo. Ver Seção 4.4;

Latent Semantic Analysis (LSA):	É uma técnica utilizada na área de Recuperação de Informação para revelar relações não explícitas entre as dimensões do vetor de características. Ver Subseção 3.10.1;
Naïve Bayes:	Técnica de classificação supervisionada, baseada no paradigma generativo, fundamentado no teorema de Bayes, que possui uma hipótese de independência entre os atributos do vetor de características. Ver Subseção 3.8.2;
SIFT:	Acrônimo do termo em inglês <i>Scale-invariant Feature Transform</i> (Transformada de Características Invariante à Escala) é uma técnica para extração de características visuais, de abordagem local, puramente espacial, proposta por Lowe (1999). Ver Subseção 3.6.2;
STIP:	Acrônimo do termo em inglês <i>Space-Time Interest Points</i> (Pontos de Interesse no Espaço-Tempo) é uma técnica para extração de características visuais, de abordagem local, espaço-temporal, proposta por Laptev (2005). Ver Subseção 3.6.2;
Support Vector Machines (SVM):	Técnica de classificação supervisionada, baseada no paradigma discriminativo, que tem por característica separar as categorias em regiões delimitadas por hiperplanos identificados com base no critério de maximização de margem. Ver Subseção 3.8.1;
Singular Value Decomposition (SVD):	É a técnica utilizada neste trabalho para realizar uma transformação geométrica no espaço de projeção das características visuais, convertendo-o para o espaço de tópicos latentes. Ver Wall et al. (2003);
Thread:	É uma coleção de vídeos, contendo um vídeo original e o conjunto de vídeos-resposta postados para ele. Ver Seção 1.2;
Vídeo Legítimo:	É um vídeo-resposta que apresenta uma relação semântica com o vídeo original, compartilhando a mesma temática em discussão. Este tipo de vídeo é considerado como sendo colaborativo. Ver Seção 1.2;
Vídeo não colaborativo:	É um vídeo-resposta cujo conteúdo não contribui positivamente com o enriquecimento da discussão na <i>thread</i> a qual está associado. Ver Seção 2.2;
Vídeo Original:	É o vídeo que origina uma <i>thread</i> de discussão em uma rede social, recebendo respostas (comentários) na forma de outros vídeos. Ver Seção 1.2;
Vídeo Spam:	É um vídeo-resposta com conteúdo normal, porém divergente à temática em discussão na <i>thread</i> . Ver Seção 1.2;
Vídeo Stuffing:	É um vídeo-resposta com pouco conteúdo próprio e,

sempre divergente da temática em discussão; são observados vários vídeos-resposta com conteúdo muito similar, geralmente utilizados para aumentar a popularidade do vídeo-original. Ver Seção 1.2;

Vídeo-referência: É a referência utilizada para comparar o conteúdo visual de todos os vídeos de uma *thread*, representando a temática discutida na *thread*. Neste trabalho o vídeo original sempre é adotado como sendo o vídeo-referência em sua *thread*. Ver Seção 6.2;

Vídeo-resposta: É um vídeo postado como comentário a outro vídeo (o vídeo original). Ver Seção 1.2.