# Segmentação Semântica de Vídeo Digital

*Danilo Barbosa Coimbra*

# Segmentação Semântica de Vídeo Digital

*Danilo Barbosa Coimbra*
mmanzato@icmc.usp.br

Orientador:
*Prof. Dr. Rudinei Goularte*
rudinei@icmc.usp.br

Monografia apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC, USP, para o Exame de Qualificação, como parte dos requisitos para a obtenção do título de Mestre em Ciências de Computação e Matemática Computacional.

USP - São Carlos
Fevereiro de 2009

COIMBRA, D. B. *Segmentação Semântica de Vídeo Digital*, São Carlos, 2009. (Monografia de Qualificação de Mestrado) - Instituto de Ciências Matemáticas e de Computação - ICMC, USP.

# Resumo

A evolução tecnológica, juntamente com o crescimento e desenvolvimento da Web, requisitam maneiras eficazes de manipulação e gerenciamento de informações. Embora a tendência das aplicações computacionais seja satisfazer do melhor modo a necessidade dos usuários, ainda existem problemas referentes à impossibilidade ou dificuldade de acesso à essas informações. Mecanismos de adaptação avançado de conteúdo visam diminuir essa dificuldade, contudo, se tratando de conteúdo multimídia o desafio é ainda maior. Uma das maneiras de adaptação de conteúdo é por meio da segmentação temporal de vídeo, a qual particiona o vídeo em partes menores. Para facilitar a recuperação e acesso ao conteúdo do vídeo, é necessário que essas partes estejam relacionadas semanticamente, por esse motivo a segmentação semântica de vídeo ou a detecção de estruturas conhecidas como cenas passam a ter mais relevância no gerenciamento de conteúdo multimídia. Desse modo, faz-se necessário o desenvolvimento de técnicas capazes de extrair o conteúdo semântico de vídeo digital, o que ainda é um questão de pesquisa em aberto. Este trabalho tem como objetivo desenvolver um método de segmentação de vídeo em cenas baseado na semântica contida no vídeo. Como resultado, espera-se que as técnicas desenvolvidas possam beneficiar diversas aplicações multimídia, tais como TV digital interativa, aprendizado eletrônico e telemedicina.

**Palavras-chave:** detecção de cenas, extração de informação, segmentação semântica, segmentação temporal, adaptação de conteúdo, personalização de conteúdo.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Lista de Siglas

ACM – *Association for Computing Machinery*

CBVR – *Content-Based Video Retrieval*

CPU – *Central Processor Unit*

DCT – *Discrete Cossin Transform*

ICMC – Instituto de Ciências Matemáticas e de Computação

IEEE – *Institute of Eletrical and Eletronics Engineers*

ISO – *International Standard Organization*

LSI – *Latent Semantic Indexing*

LSA – *Latent Semantic Analysis*

MPEG – *Moving Picture Experts Group*

NCut – *Normalized Cut*

OWL – *Ontology Web Language*

RDF – *Resource Description Framework*

SVM – *Support Vector Machine*

TV – Televisão

USP – Universidade de São Paulo

UMA – *Universal Multimedia Access*

WWW – *World Wide Web*

Capítulo

1

# Introdução

A transmissão de informações audiovisuais é a principal diferença entre a televisão e outros meios de comunicação como rádio, jornais ou revistas, além de ser também a razão de seu sucesso como principal mecanismo de comunicação durante décadas, todavia novos tempos trazem novos usuários, os quais criam novas necessidades. A contínua evolução digital, juntamente com a expansão da WWW (World Wide Web), estão ocasionando transformações tanto no conteúdo quanto na comunicação dessas informações audiovisuais. Os vídeos analógicos que outrora eram transmitidos pela televisão ou reproduzidos em fitas cassetes, agora são transmitidos ou reproduzidos como vídeos digitais em diversas aplicações. Telemedicina, aprendizado eletrônico, bibliotecas digitais, videoconferência, TV Digital e, por consequência, a TV Interativa, são todos exemplos de aplicações que utilizam conteúdo multimídia, especificamente o vídeo digital.

A TV Interativa, em especial, utiliza a tecnologia digital para fornecer serviços interativos aos usuários, de modo que acessem o conteúdo multimídia da TV da mesma maneira que interagem com os programas da Web. Assim, a convergência entre TV, multimídia e Web tem estimulado o desenvolvimento de aplicações que possuem como objetivo oferecer serviços personalizados, interação e busca baseado em objetos, entrega e recepção independentes do dispositivo, entre outros (GOULARTE, 2003).

Em paralelo a busca por interatividade, os usuários fazem uso de diferentes tipos de dispositivos computacionais, incluindo dispositivos móveis, para ter acesso ao conteúdo multimídia, o qual é disponibilizado tanto pela Web quanto pela recente transmissão de sinal digital de televisão. Contudo, algumas limitações podem dificultar o acesso aos dados, como uso de dispositivos com características limitadas, oscilações consideráveis na largura de banda e preferências adversas do usuário.

Uma das alternativas pesquisadas atualmente para possibilitar o acesso transparente

ao conteúdo multimídia, é a personalização e adaptação de conteúdo. Enquanto a adaptação decide a melhor versão de conteúdo para ser apresentado e a melhor maneira de fazê-lo (LUM; LAU, 2002), a personalização é vista como um caso particular da adaptação, pois os dados são adaptados conforme as necessidades do usuário (BARRIOS; MöDRITSCHER; GüTL, 2005). Magalhães e Pereira (2004) também apresentam a adaptação de conteúdo com preferências do usuário como uma operação de customização do conteúdo multimídia.

Um dos maiores desafios para viabilizar o acesso a conteúdo personalizado está em realizar a manipulação de vídeos digitais a fim de torná-lo customizável. O primeiro passo para gerenciar tal conteúdo é realizar a indexação, gerando sumários ou índices que facilitem seu acesso. No caso de documentos de texto os índices servem como um elemento que auxilia a encontrar uma determinada informação nas diferentes seções ou capítulos do documento (SRINIVASAN; NEPAL, 2005). Entretanto, criar um índice similar que aponta para diferentes partes de um vídeo é uma tarefa complexa. A tentativa manual de indexar o conteúdo audiovisual pode ser subjetiva, uma vez que existem diferentes maneiras de descrever conteúdo multimídia dependendo do usuário, do propósito do uso e da tarefa que precisa ser realizada (SRINIVASAN; NEPAL, 2005). Portanto, nota-se que a recuperação de informação em vídeos digitais não é uma tarefa trivial.

Esforços têm sido realizados em direção à extração de informações do conteúdo de vídeo digital, em especial, pela área de Análise de Vídeo Digital Baseado em Conteúdo (do inglês, *Content Based Analysis of Digital Video*). Seu propósito está em fornecer subsídios para o desenvolvimento de sistemas que tornam o vídeo mais "compreensível" por meio de algoritmos e técnicas que empregam, ou não, o uso de semântica. Áreas relacionadas a computação como processamento digital de imagem e som, reconhecimento de fala, recuperação de informação em texto, processamento de linguagem natural, reconhecimento de padrões, entre outras, compõem a Análise de Conteúdo de Vídeo, tornando evidente a sua posição como uma área de pesquisa multidisciplinar (HANJALIC, 2004). O foco dessa área de pesquisa está em fornecer modos de interação com o vídeo por meio de técnicas de indexação e navegação e/ou representação, possibilitando o seu uso em aplicações que realizam a entrega do conteúdo do vídeo personalizado ao usuário (HANJALIC, 2004). Entretanto, anteriormente aos passos de indexação e navegação, é necessário efetuar a segmentação temporal do fluxo do vídeo.

A segmentação temporal é considerada a primeira etapa no processo de extração de informação desse conteúdo, além de ser responsável por identificar segmentos com significado. Na literatura, as técnicas relacionadas à este tipo de segmentação trabalham com conceitos similares sobre a definição da estrutura de um vídeo. Geralmente esta estrutura divide-se em quatro partes: quadros, tomadas, cenas e o vídeo completo. Visto que o quadro é a menor unidade do vídeo, além de ser uma imagem estática, a segmentação da próxima estrutura (tomada) é o início do processo de extração de informação. Enquanto

a tomada é definida como um conjunto de quadros gravados da operação de uma única câmera, a cena é caracterizada como uma coleção de tomadas consecutivas e relacionadas umas com as outras por meio de conteúdo semântico (ANER-WOLF; KENDER, 2004).

A definição de informação semântica não é trivial, pois depende da interpretação de cada pessoa e do tipo de conteúdo que está associado. O processo de segmentação incluindo semântica é um modo de sumarizar todo o conteúdo do vídeo digital em segmentos com significados semelhantes e compreensíveis ao usuário (DONG; LI, 2006). Tanto a detecção de alguns elementos espaciais (objetos) quanto a detecção de eventos específicos nos vídeos (gols em um jogo de futebol, notícia de clima em telejornal) compõem a segmentação semântica. Entretanto, o modo que o usuário relembra de vídeos é associada a eventos ou histórias (WANG; CHUA, 2003), por isso a necessidade de representar o conteúdo em segmentos semânticos similares correlacionados temporalmente, ou seja, realizar a segmentação semântica de vídeo digital.

Características como quantidade de mídias, formatos de compressão e diversos gêneros de vídeos fazem parte da metodologia que compõem os trabalhos relacionados ao tema. O estado da arte é obtido utilizando a maior parte de informações providas pelo vídeo, isto é, usando técnicas que incluem dados de mais de uma mídia (abordagens multimodais), as quais tendem a aumentar a eficiência da extração da semântica do conteúdo (SNOEK; WORRING; SMEULDERS, 2005; WANG et al., 2008). Nessas técnicas, inclusive, as informações obtidas de texto, como legendas ou *closed-captions*, podem auxiliar a classificação das cenas (PAO et al., 2008; MANZATO; GOULARTE, 2008). Contudo, anotações não foram exploradas na literatura. Ainda, algoritmos que empregam técnicas de aprendizagem de máquina são frequentemente adotados para a recuperação de informações semântica (LIU; HE; ZHANG, 2007; JIN; SHI; CHUA, 2004; TOWN; SINCLAIR, 2001).

Desse modo, este plano tem o objetivo de desenvolver um método de segmentação de vídeo em cenas baseado na semântica contida no vídeo, explorando as vantagens identificadas nas técnicas da área de Análise de Conteúdo de Vídeo. Um dos desafios a ser investigado é a lacuna entre a interpretação do usuário para o conteúdo de imagens e a interpretação que o computador possui ao extrair os dados dessas imagens, também chamada de lacuna semântica.

Este trabalho está organizado do seguinte modo. No capítulo 2, Conceitos e tecnologias, são apresentados os principais conceitos relacionados ao estudo do vídeo digital. No capítulo 3, Revisão Sistemática, realiza-se um levantamento bibliográfico sistematizado em torno dos trabalhos desenvolvidos com o tema da pesquisa. Os trabalhos relacionados são descritos no capítulo 4, juntamente com suas metodologias e técnicas no processo de segmentação de vídeo, assim como os desafios da área e seu estado da arte. No capítulo 5, Plano de Pesquisa, apresenta-se o projeto de pesquisa a ser desenvolvido em conjunto com um esboço de uma técnica como contribuição ao tema relacionado.

Capítulo

2

# Conceitos, Definições e Tecnologias

## 2.1 Considerações Iniciais

O vídeo, formado por sinais analógicos ou digitais, tem como funções a captura, armazenamento, transmissão ou apresentação de imagens em movimento. Seu uso principal, a partir de meados do século XX, é concebido pelo principal meio de comunicação ainda hoje, a televisão. O princípio de funcionamento das primeiras versões desse aparelho era baseado em transmissões analógicas, as quais utilizavam ondas eletromagnéticas contínuas. Todavia, com a evolução tecnológica, os vídeos passaram a ser disponibilizados em outro tipo de formato, o formato digital.

O vídeo digital é formado por um sinal com valores discretos (descontínuos) no tempo e em amplitude, definido para determinados instantes de tempo e assumindo um conjunto de valores finito. Assim, as funções definidas anteriormente para o vídeo são mais eficientes e eficazes quando utilizado esse tipo de formato. A sua maior aplicabilidade encontra-se em equipamentos computacionais e com tendência de substituição dos televisores analógicos por modelos digitais, ocasionando o aumento da qualidade visual e possibilidades de interação com o conteúdo assistido (TV Interativa).

A popularidade do uso de vídeos digitais são consequências de fatores que estão ocorrendo concomitantemente: avanços na tecnologia de compressão, maior acesso às câmeras digitais, dispositivos e sistemas com alta capacidade de armazenamento, aumento do uso da Internet e redes banda larga (HANJALIC, 2004). Esses fatores ocasionam maior demanda por aplicações que fazem uso de vídeo como um importante mecanismo de transmissão de informações audiovisuais. Exemplos dessas aplicações podem ser encontradas, inclusive, em áreas sociais como saúde, educação e segurança, respectivamente com aplicabilidades específicas na telemedicina, educação à distância por meio do aprendizado

eletrônico, em sistemas de monitoramento e vigilância, entre outras.

Além de aplicações em áreas sociais, outros cenários são beneficiados em sua utilização. Áreas como publicidade e propaganda e *marketing* fazem uso cada vez maior desse conteúdo para repassar ao cliente a necessidade da obtenção de um determinado produto. Na Internet, sítios como o Youtube [1] armazenam uma vasta quantidade de vídeos caseiros, ganhando milhões de usuários ao redor do mundo. Entretanto, outras áreas sofrem problemas com o aumento dessa popularidade, como ocorre com a indústria cinematográfica, a qual está sofrendo grandes perdas com a difusão sem autorização desse conteúdo, principalmente quando é feito o comércio ilegal de produtos com direitos autorais, ocasionando o problema da pirataria de vídeos.

A grande diversidade de aplicações denotam a importância desse conteúdo multimídia como um poderoso meio de comunicação. A quantidade de vídeos que são produzidos, assistidos, editados, armazenados, transmitidos e trocados entre usuários ocorre devido à possibilidade de realizar o processamento de vídeos de maneira automática. Entretanto, se o objetivo do usuário é, por exemplo, localizar um determinado segmento de vídeo de seu interesse em uma coleção de arquivos desse tipo, a única maneira é assistir a cada segmento desde o começo utilizando operações de avanço rápido (*fast-forward*) e retrocesso rápido (*fast-backward*) até que ele encontre o segmento desejado (ZHU; LIU, 2008a). Por meio desse exemplo fica evidente que a procura linear tem baixa eficiência e consome muito tempo, tornando o processo de busca inadequado, requisitando mecanismos de recuperação de informação mais rápidos e melhores.

Portanto, a função que explora a extração dos dados do vídeo com o objetivo de obter informações sobre seu conteúdo é um problema de pesquisa na literatura, tornando um desafio nas áreas que realizam análises em vídeo baseados em conteúdo. Recuperação de Vídeo Baseado em Conteúdo (do inglês, *Content-Based Video Retrieval* - CBVR) e Análise de Conteúdo de Vídeo (do inglês, *Video Content Analysis*) são áreas que tentam contornar e/ou diminuir este problema. A Figura 2.1 (HANJALIC, 2004) ilustra como as áreas citadas abordam a recuperação de informação em vídeo. Na etapa inicial, são desenvolvidas técnicas ou algoritmos que fazem o processo de extração de informação do material. Após, estes são aplicados no vídeo com o intuito de detectar trechos que contenham pessoas, objetos ou eventos para que possam ser identificados e geralmente representados como imagens no dispositivo computacional. Por fim, ao escolher e selecionar uma imagem, o usuário visualiza o segmento do vídeo associado ao trecho identificado, extraído do vídeo original.

[1] http://www.youtube.com


Figura 2.1: Algoritmos aplicados nos vídeos digitais que são desenvolvidos para facilitar a recuperação de seus conteúdos (HANJALIC, 2004)

## 2.2 Estrutura do vídeo digital

O processo de extração de informação nos vídeos digitais tem como foco auxiliar o usuário no acesso a seu conteúdo. Contudo, atualmente, o modo tradicional de acesso ocorre de maneira linear, sendo necessário que o usuário procure o segmento desejado desde o começo. Para que o acesso seja efetuado de modo não linear é necessário entender como é constituído este tipo de componente.

Também chamada de representação hierárquica (LI; MING; KUO, 2001; RUI; HUANG; MEHROTRA, 1998), a estrutura do fluxo do vídeo digital é constituída de níveis ou camadas, as quais são formadas por intermédio da análise de seu conteúdo (Figura 2.2). Os algoritmos de extração de dados atuam em uma ou mais dessas camadas fornecendo, algumas vezes, informações para outros algoritmos desenvolvidos nas camadas superiores. Particularmente, os algoritmos da área de Análise de Vídeo Baseado em Conteúdo estruturam o conteúdo do vídeo seguindo essa abordagem (NGO; ZHANG; PONG, 2001).

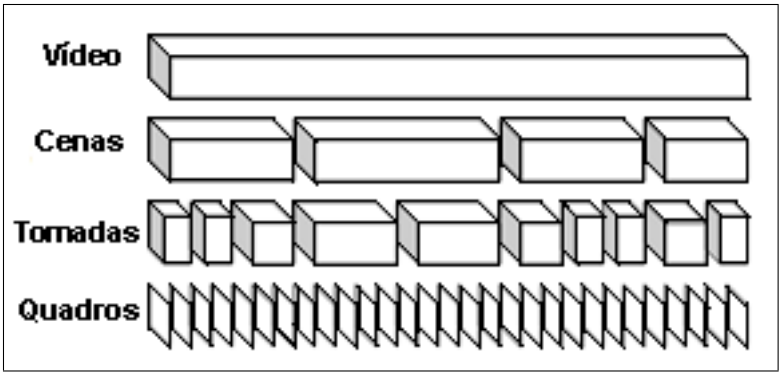
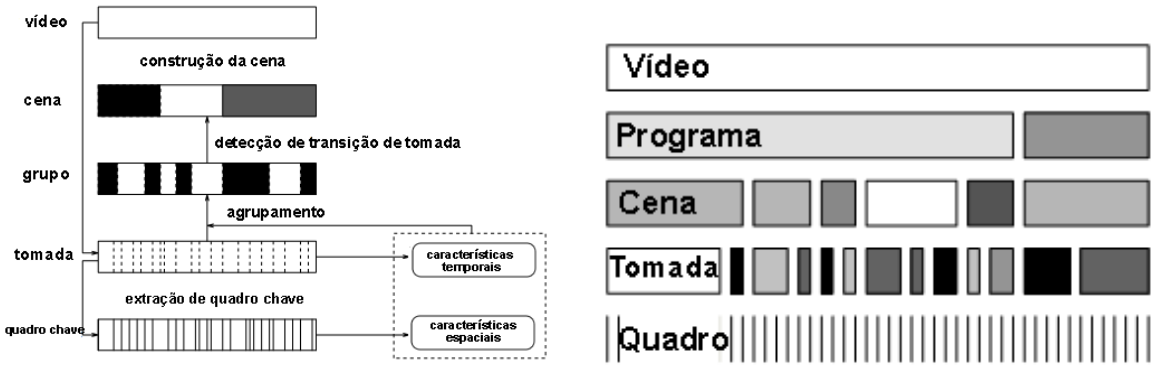



Figura 2.2: Estrutura do fluxo de vídeo digital

Sendo o vídeo uma sequência de imagens estáticas, essas imagens estão presentes na

estrutura e são chamadas de quadros. As tomadas são os seguimentos da camada acima, constituindo uma sequência de quadros gravados continuamente e representando uma ação contínua em tempo e espaço. As cenas são definidas como um grupo de tomadas com conteúdo correlacionado (ZHAO; YANG; FENG, 2001), também chamadas de unidades semânticas. A última camada é o próprio vídeo, ou vídeo em sua forma "crua" (do inglês, *raw video*). Basicamente, a diferença entre as camadas de tomadas e cenas está na natureza da análise, pois quanto mais baixo no nível da estrutura, maior a eficácia das técnicas e menor a complexidade computacional. Entretanto, quanto mais alto no nível, maior a dependência do gênero do vídeo (e.g. noticiário, evento esportivo, filmes de ação, etc) (ZHANG, 2006).

Embora a representação da estrutura de vídeo (Figura 2.2) seja considerada um consenso na comunidade acadêmica (SURAL; MOHAN; MAJUMDAR, 2005; OH et al., 2005; ZHAO; YANG; FENG, 2001), alguns autores utilizam outras representações (Figura 2.3). Mesmo não sendo iguais, há muita semelhança entre todas as apresentadas, havendo mudanças somente no acréscimo de uma camada na estrutura, seja entre as camadas de tomadas e cenas ou entre as camadas de cenas e o vídeo completo. No primeiro caso, é apresentada uma estrutura com uma camada Grupo que representa uma etapa responsável pelo agrupamento de segmentos de tomadas (RUI; HUANG; MEHROTRA, 1998)(Figura 2.3(a)). No segundo caso, é formada uma camada Programa para representar possíveis episódios de uma série de TV (camada Vídeo) (AL-HAMES et al., 2006) (Figura 2.3(b)). Em ambos fica evidente que o tipo de estrutura é elaborada de acordo com a metodologia sugerida, dependendo do nível de granulação de informação que o autor pretende trabalhar.




(a) (RUI; HUANG; MEHROTRA, 1998) (b) (AL-HAMES et al., 2006)

Figura 2.3: Outras representações para estrutura de vídeo digital

Apesar do processo de análise de vídeo ser realizado de modo qualitativa, buscando resultados eficazes, Al-Hames et al. (2006) faz uma análise quantitativa da estrutura, proporcionando uma idéia da quantidade de cada estrutura e da complexidade de trabalhar com vídeos digitais. As estimativas apresentadas nessa análise compreendem uma aproximação de vinte e cinco quadros por segundo, muitas centenas de tomadas por hora e

cerca de cem cenas por hora.

## 2.3 Análise do Vídeo Digital

A identificação de tais estruturas é uma etapa essencial na criação de mecanismos que viabilizam o acesso ao teor do material disponível no vídeo, porém constitui apenas uma etapa no gerenciamento desse conteúdo multimídia. A área de Análise de Vídeo Digital Baseado em Conteúdo é composta por três grupos distintos, cada qual contendo algoritmos que realizam análise no conteúdo de vídeo (HANJALIC, 2004):

- Análise da estrutura do vídeo: relacionada à segmentação de vídeo em estruturas temporais de baixo nível (tomadas) ou alto nível (cenas).
- Indexação de conteúdo de vídeo: designa automaticamente pedaços dos dados do vídeo para categorias pré-especificadas no formato de rótulos (*e.g.* alegria, futebol no estádio, criança chorando, etc). A ligação para essas categorias ocorre por meio de elementos denominados índices, possibilitando a recuperação posterior de cada segmento por meio desses elementos.
- Representação e abstração do conteúdo de vídeo: constrói resumos compactos mas compreensíveis dos segmentos de vídeo. O objetivo dessa etapa é comunicar de maneira eficiente e eficaz o conteúdo do vídeo para o usuário.

De modo resumida, a análise da estrutura extrai do vídeo estruturas temporais menores (tomadas ou cenas) (DIMITROVA et al., 2002). A indexação do conteúdo de vídeo designa esses segmentos para um grupo de categorias, formando índices que relacionam os segmentos. Por fim, na representação e abstração de conteúdo, os índices são disponibilizados e apresentados aos usuários de maneira que estes possam navegar no conteúdo do vídeo. Disponibilizar esse tipo de conteúdo multimídia visando modos de busca e interação com o usuário também é uma maneira de realizar adaptação e personalização de conteúdo (LUM; LAU, 2002; BARRIOS; MöDRITSCHER; GüTL, 2005).

O método de segmentar um vídeo constitui da extração de partes relevantes do vídeo. A segmentação pode ser classificada como espacial ou temporal (MAGALHãES; PEREIRA, 2004). A segmentação espacial foca a divisão do vídeo baseada em determinadas características espaciais (objetos), detecção de faces e reconhecimento de objetos que não estejam no plano de fundo representam algumas de suas técnicas. Na segmentação temporal são classificados segmentos com menor duração de tempo e com características semelhantes, eventos como o gol em um vídeo de futebol ou explosões num filme de ação são exemplos dessa segmentação. O foco desse trabalho está em recuperar informações baseadas em eventos, visto que o usuário assiste e recorda de vídeos em termos de eventos, episódios ou histórias (WANG; CHUA, 2003).

Wang e Chua (2003) descrevem a segmentação temporal do vídeo como sendo de dois tipos: a sintática e a semântica. A segmentação sintática ocorre quando a técnica ou algoritmo é empregado para identificar tomadas, visto que sua análise pode ser efetuada extraindo dados dos quadros que a compõem. Na segmentação semântica é proposta a identificação de cenas, as quais requerem um melhor entendimento dos tópicos presentes num determinado conjunto de tomadas. A elaboração das metodologias nos algoritmos ou técnicas usados para realizar a etapa de análise de estrutura do vídeo possui algumas variações. A seguir, são descritas as abordagens consideradas nesse processo (HANJALIC, 2004; NGO; ZHANG; PONG, 2001):

- Domínio com compressão X domínio sem compressão: o processamento de vídeo digital demanda muito tempo computacional e uma grande quantidade de memória. Para que ambos, tempo e espaço, sejam reduzidos, abordagens que manipulam vídeos diretamente em formato comprimido tem se tornado uma pratica constante.

- Técnica automática X técnica semi-automática: para alcançar o melhor nível de interação com o usuário, a aplicação deve requisitar somente o nível de interação necessário. Enquanto sistemas de vigilância devem funcionar de modo automático, detectando movimentos suspeitos e avisando ao usuário, outros sistemas fornecem liberdade ao usuário para determinar, por exemplo, o tamanho dos índices de um certo filme ou os melhores lances de um esporte.

- Uso de várias mídias: além da parte visual, o vídeo pode ser acompanhado de outras mídias como áudio e texto. O fluxo de áudio pode consistir de música, representar vozes (fala), som ambiente ou ainda uma mistura dos três. Os textos geralmente representam a tradução de uma língua estrangeira, disponibilizados em formatos de legendas, ou descrevendo qualquer som presente no vídeo (palmas, passos, risos, músicas, fala, etc), por meio do *closed caption*. Caso o conteúdo do vídeo seja composto por mais de um tipo de mídia, o ideal é que as informações de cada mídia sejam agrupadas de maneira que proporcione maior significado ao conteúdo, ou seja, maior semântica ao vídeo.

Outra abordagem discutida é em relação ao momento em que as técnicas são empregadas, pois a transmissão do vídeo pode ser efetuada em tempo real (CORREIA; PEREIRA, 2004). Portanto, uma nova divisão das abordagens faz-se necessária: as que realizam processamento em tempo real ou as que realizam em vídeos já armazenados, também conhecida como *off-line*.

Nas sub-seções a seguir são detalhadas as estruturas temporais que compõem o fluxo de vídeo digital.

### 2.3.1 Detecção de Tomadas

O desenvolvimento de algoritmos nessa área tem a maior e mais rica história na área de análise de conteúdo de vídeo. Maior porque é a área que iniciou, de fato, as tentativas de detecção automática de cortes em vídeos, e mais rica porque contem a maioria dos trabalhos publicados na área desde então (HANJALIC, 2004). Os trabalhos que abordam a detecção de tomadas ou detecção de transição de tomadas fornecem a base para quase todas as abordagens de análise de conteúdo de vídeo de alto nível (cenas), além de ser também um pré-requisito para o desenvolvimento da estrutura do conteúdo do vídeo. As definições sobre essa estrutura variam, Koprinska e Carrato (2001) definem a tomada como uma sequência linear de quadros retirados de uma única câmera, enquanto Dimitrova et al. (2002) identifica o limiar da tomada como pontos de edição ou pontos em que ocorrem ligamento/desligamento da câmera. Por meio dessas definições nota-se que tomadas são segmentos compostos por quadros sendo dependente das ações de câmeras.

Entre duas tomadas consecutivas ocorre uma transição, a qual é separada em dois grupos: abrupta ou gradual. Também chamada de corte, a transição abrupta é mais fácil de ser detectada, pois consiste de uma mudança instantânea entre uma tomada e outra, ocorrendo como um corte entre dois quadros consecutivos (quadros 2 e 3) (Figura 2.4) (KOPRINSKA; CARRATO, 2001).


Figura 2.4: Transição abrupta de tomada

A transição gradual de tomadas é mais difícil de ser realizada e pode ser dividida em duas classes: aquelas que ocorrem simultaneamente, mas que afetam gradualmente todo o pixel da imagem e àquelas que afetam abruptamente todo um conjunto de pixels, com este conjunto mudando em cada quadro (JOYCE; LIU, 2006). Transições como *fade in/out* e dissolução fazem parte do primeiro grupo e *wipes* constituem o segundo, todas essas descritas e ilustradas a seguir:

- Dissolução: quando uma imagem sobrepõe outra de modo gradativa, isto é, estendendo em vários quadros (Figura 2.5).
- *Fade in* e *Fade out*: quando uma imagem clara escurece gradativamente ou o oposto, uma imagem escura clareia gradativamente (Figura 2.6). Esse pode ser considerado um caso especial de dissolução, uma vez que ocorre gradativa sobreposição de quadros claros à escuros ou o oposto (NGO; ZHANG; PONG, 2001).

Figura 2.5: Dissolução (PORTER; MIRMEHDI; THOMAS, 2003)

Figura 2.6: Transição *fade out* seguida por *fade in* (KOPRINSKA; CARRATO, 2001)

- *Wipes*: quando uma imagem é "empurrada", "dobrada" de alguma modo ou direção, até desaparecer, dando lugar a uma outra imagem. A Figura 2.7 representa quatro diferentes tipos de *wipes*.


Figura 2.7: Diferentes tipos de *wipes* (JOYCE; LIU, 2006)

Devido a maior parte do processamento computacional ser dedicada a esses algorit-

mos de detecção de transição de tomadas, o seu nível de complexidade deve ser baixo. Entretanto, minimizando o nível de complexidade acarreta numa menor taxa de detecção de erros. Um exemplo é o problema da transição gradual de tomadas os quais os efeitos editáveis estão sobrepostas em movimentos de objetos e câmera. Para eliminar a influência da movimentação no comportamento do sinal entre a transição, estimativas e compensação de movimentação podem ser aplicados. Contudo, isso é computacionalmente caro, justificável apenas em casos onde a informação de movimento está realmente disponível, como em vídeos comprimidos (e.g. MPEG), tornando o desempenho do detector dependente das características particulares de codificação (HANJALIC, 2004).

Quando segmentando um vídeo em tomadas, dois problemas devem ser considerados. O primeiro é a capacidade de distinguir entre uma transição de tomada e uma mudança ocorrida dentro de uma tomada. A maioria das mudanças normais são de movimentos de objetos ou movimento de câmera. Durante esses movimentos, o conteúdo das imagens podem alterar drasticamente, como por exemplo, a movimentação de objetos grandes ou rápida movimentação da câmera torna difícil a identificação de transição de tomadas. O segundo problema é a capacidade de distinguir entre transições graduais e quadros sem transição. Quando transições graduais estão envolvidas, duas tomadas estão mescladas no processo de edição: a evolução de uma tomada à outra ocorre ao longo de vários quadros, sendo cada um diferente do outro apenas por pequenos detalhes. Usando o método de detecção normal de corte (abrupto), esse efeito especial pode não ser detectado como uma mudança de tomada.Portanto, é necessário o uso de algoritmos dedicados para transição gradual de tomadas.

### 2.3.2 Detecção de Cenas e a Lacuna Semântica

Enquanto a detecção de tomadas é o primeiro passo para a realização da análise do vídeo, a detecção de cenas é o primeiro passo em direção a compreensão semântica do vídeo digital (CHEN; LAI; LIAO, 2008). Seguindo o fluxo contrário dessa definição, nota-se que a compreensão semântica depende de estruturas denominadas cenas. Segundo Aner-Wolf e Kender (2004), cena é uma coleção de tomadas consecutivas que estão relacionadas umas com as outras por meio de conteúdo semântico. A expressão "semântica" também está presente na definição de Zhai e Shah (2006), afirmando que uma cena é um grupo de tomadas relacionadas semanticamente e coerente de acordo com um tema ou assunto. Com base nessas definições é possível identificar algumas características desse segmento: *i)* são compostos por um grupo de tomadas; *ii)* essas tomadas devem conter um tema ou assunto semelhante entre si; *iii)* informação semântica deve estar presente.

De acordo com o dicionário Houaiss (HOUAISS, 2001) a palavra semântica é apresentada como o estudo do significado das palavras, contudo, a unidade relevante na área de segmentação de vídeo não são palavras, mas sim segmentos de vídeo. Assim, a semântica

relacionada ao vídeo denota-se ao seu próprio significado ou de um seguimento associado ao mesmo. Portanto, a segmentação semântica de vídeo, ou extração de cenas, está relacionada com a extração de unidades que tenham significado similares (semântica), de acordo com um determinado tema ou assunto e decorrente de um agrupamento de tomadas (SURAL; MOHAN; MAJUMDAR, 2005).

Contudo, determinar o significado de uma cena não é uma tarefa simples. A distância entre a informação que pode ser extraída do conteúdo visual e a interpretação ou significado desses dados por um usuário em determinada situação é visto como uma questão em aberto, também conhecida como lacuna semântica (SMEULDERS et al., 2000). As informações extraídas do vídeo, geralmente de estruturas como quadros e tomadas, são chamadas de características de baixo-nível, e são representadas por dados como cor, forma, textura, etc., possibilitando o uso de diversas técnicas de extração (Tabela 2.1).

Tabela 2.1: Tabela de algumas características de baixo-nível com suas respectivas técnicas (HANJALIC, 2004)

| **Características** | **Técnicas** |
|---|---|
| cor | distribuição de cor, momentos de cor |
| textura | energia de textura, contraste, repetividade, complexidade, modelos estocásticos, modelo auto-regressivo, auto-correlação |
| forma | estatísticas de borda, parâmetros de curvatura |
| áudio | *pitch*, espectro de frequência, características de sinais temporais, fonemas, *zero-crossing rate* |
| movimentação | direção e intensidade do movimento, coerência de campo de movimentação |
| relacional | relações direcionais e topológicas entre linhas, regiões ou objetos |

A interpretação que o usuário fornece de determinado segmento do vídeo são as informações de alto nível, as quais encontradas entre as transições de seguimentos semânticos (HANJALIC, 2004). Tais informações representam as cenas, ou eventos, e podem ser encontradas nos vídeos dos mais variados gêneros: um evento em vídeo de esporte (como um gol no futebol e uma cesta no basquete), determinada notícia em um telejornal (como notícia de política, economia, clima, etc.), cenas em filmes de ação (explosões e corridas de carro).

Desse modo, o espaço entre as características de baixo nível e as de alto nível representa a lacuna semântica (ilustrada na Figura 2.8). Essa lacuna é a responsável por se considerar a detecção de tomadas como sintática e o detecção de cenas como semântica. Enquanto que para detectar tomadas, as características de baixo nível são suficientes, na detecção de cenas é necessário obter as características de alto nível (WANG; CHUA, 2003).

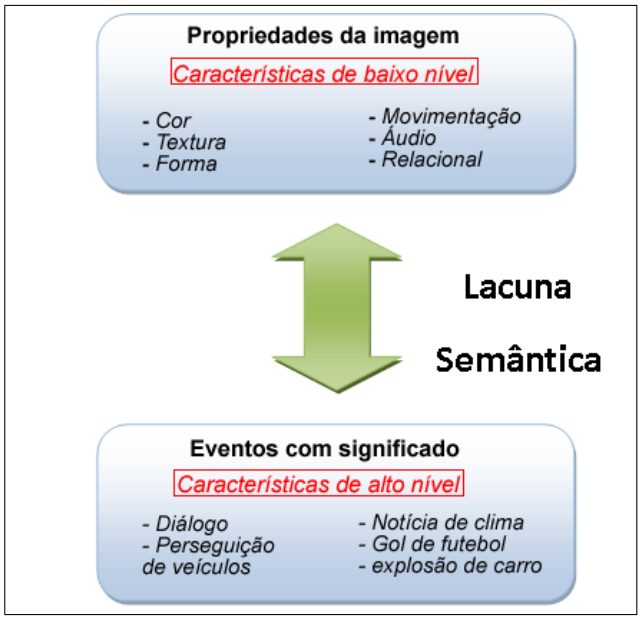



Figura 2.8: Representação da lacuna semântica. Adaptado de (HANJALIC, 2004)

A Figura 2.9 apresenta uma relação dos segmentos, representados por tomadas e cenas, e suas respectivas características associadas, baixo e alto nível, respectivamente.

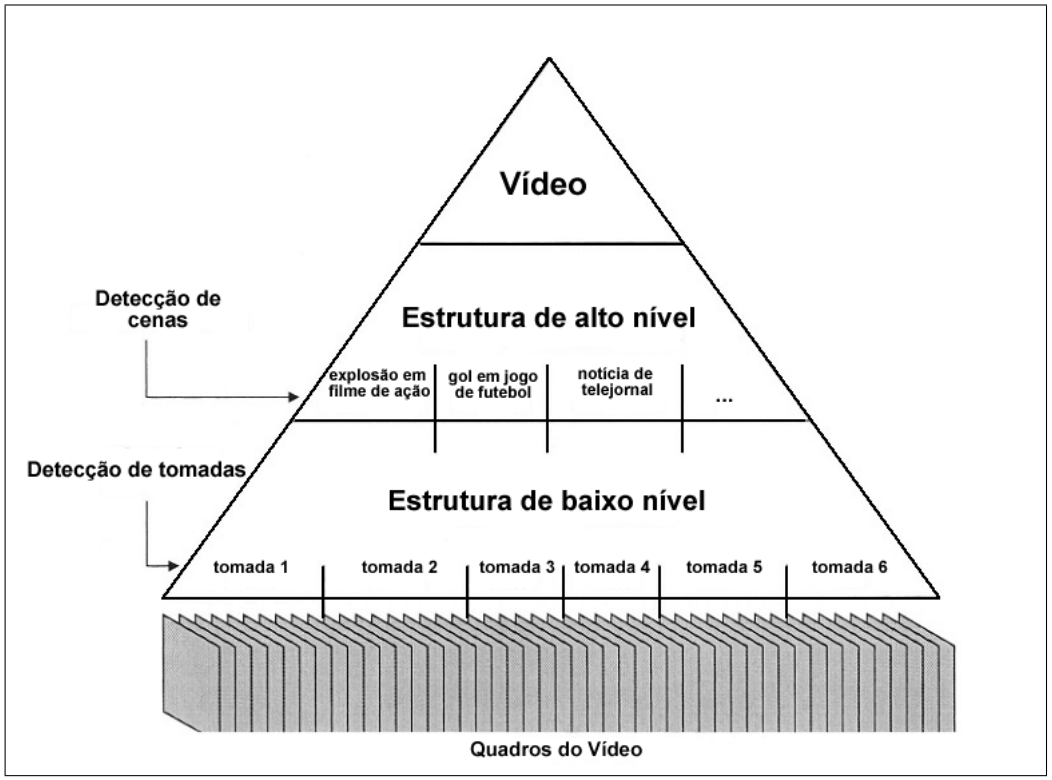



Figura 2.9: Pirâmide da estrutura do conteúdo do vídeo. Adaptado de (HANJALIC, 2004)

As aplicações relacionadas a segmentação/identificação de cenas ocorre em vários domínios, fornecendo diversos benefícios: em filmes menores a segmentação de cenas prove capítulos que correspondem a diferentes subtemas do filme; em vídeos de tele-

visão, a segmentação pode ser usada para separar os comerciais dos programas comuns. Nos noticiários de televisão, a segmentação pode ser utilizada para identificar diferentes histórias jornalísticas (tal como clima, economia, politica, esportes, etc). Em vídeos caseiros, pode ajudar os usuários a organizar logicamente os vídeos relacionados a eventos distintos (aniversários, formatura, casamento, férias, etc.) (ZHAI; SHAH, 2006).

## 2.4 Avaliação de Resultados

O modo de avaliação dos resultados na segmentação temporal do vídeo, tanto na detecção de cena quanto na detecção de tomadas, frequentemente ocorre por meio de medidas de avaliação quantitativas, denominadas *precision* e *recall*. Não somente encontradas na segmentação de vídeo, essas medidas são comumente utilizadas na avaliação de desempenho dos algorítmos de recuperação de informação (BAEZA-YATES; RIBEIRO-NETO, 1999). O intuito de tal método é verificar a eficiência com base nos segmentos detectados de maneira correta (*Precision* 2.1) e também avaliando casos em que detectou-se transições onde não ocorreram (*Recall* 2.2). As medidas são descritas a seguir:

$$precision = \frac{num.\ verdadeiro\ positivos}{num.\ verdadeiro\ positivos\ +\ num.\ falso\ positivos} \tag{2.1}$$

$$recall = \frac{num.\ verdadeiro\ positivos}{num.\ verdadeiro\ positivos\ +\ num.\ falso\ negativos}, onde \tag{2.2}$$

num.verdadeiro positivos é a quantidade de segmentos corretos que foram detectadas pela técnica; num. falso positivo é a quantidade de segmentos que foram detectadas mas que não são corretos, ou seja, não existe e num. falso negativo é a quantidade de segmentos que não foram detectados mas que existem.

Apesar dessa avaliação ser encontrada quase que na totalidade dos trabalhos de segmentação, algumas limitações são identificadas. Baeza-Yates e Ribeiro-Neto (1999) relatam que uma estimativa apropriada para o *recall* requer um conhecimento detalhado de todos os documentos envolvidos da amostra, principalmente se o conjunto da amostra for muito grande. O fato das duas medidas capturarem diferentes aspectos do conjunto de documentos analisados pode ser interpretado como outro problema, possivelmente resolvido com uma abordagem que combine essas duas medidas (e.g. O significado Harmônico [2] e a Medida E [3]).

Zutschi et al. (2005) também descreve algumas limitações sobre o uso dessa avaliação

[2]Composta por uma função que obtém como resultado um valor entre 0 e 1. Sendo 0 quando não houve sucesso na recuperação de informação ou 1 quando ocorreu sucesso.

[3]Por meio de uma constante adicionada à formula, o usuário pode definir qual das duas medidas é mais importante, funcionando como um peso.

ao relatar que a utilidade prática dessas medidas estão sendo questionadas sob o ponto de visto do usuário final. Notadamente, a medida de precisão não é um bom indicador para representar a percepção do usuário na qualidade em sistemas de recuperação de imagens baseada em conteúdo. Corroborando as palavras de Baeza-Yates & Ribeiro-Neto, o *recall* nem sequer fez parte da avaliação por não envolver reconhecimento avançado do conteúdo da base de dados de imagens. Outra questão envolvendo a recuperação multimídia baseado em conteúdo diz respeito à natureza da relevância dos dados. Os algoritmos tendem a agrupar dados similares, no entanto, o conceito de similaridade é subjetivo. O significado, ou semântica, de um dado depende do ponto de vista de cada usuário, assim faz-se necessário explicitar melhor semântica das informações para o usuário, a fim do sistema retornar resultados mais eficientes.

Uma boa alternativa para substituir ou complementar a abordagem do *precision* e *recall* está em adotar medidas de avaliações centradas no usuário. Tais medidas fazem uso do usuário nos processos de classificação dos dados e avaliação dos resultados, possibilitando resultados mais eficientes, podendo conter um índice de significado semântico maior (ZUTSCHI et al., 2005).

## 2.5 Compressão de Vídeo Digital

Como discutido, a organização eficiente dos dados de um vídeo digital permanece como uma questão em aberto, principalmente se provêm de inúmeras origens. Consequentemente, a consistência dos dados deve ser efetuada de maneira apropriada no sentido de armazenar em um formato padrão a fim de auxiliar o acesso e recuperação. Em um contexto onde há uma grande quantidade de vídeo, ou seja, um grande volume de dados, a redução do espaço de armazenamento torna-se desejável, necessitando a criação de um formato de compressão.

Foi com o intuito inicial de comprimir o espaço de dados digitais e, posteriormente, promover a interoperabilidade entre diversas aplicações multimídia, que o grupo MPEG (*The Moving Picture Coding Experts Group* - Grupo de Especialistas de Codificação de Imagens em Movimento) criou, em 1988, uma família de padrões (MPEG-1, MPEG-2, MPEG-4, MPEG-7 e MPEG-21). Como um grupo desenvolvido pela ISO (*International Organisation for Standardisation* - Organização Internacional de Padrões), este tem como finalidade desenvolver padrões internacionais de compressão, descompressão, processamento, codificação e representação de imagens em movimento, áudio e suas combinações, com o intuito de satisfazer a maior variedade de aplicações (WU; IRWIN; DAI, 2001).

Dentre as práticas mais adotadas no processo de análise do vídeo está o uso desses padrões de compressão. MPEG-1, MPEG-2 e MPEG-4 são os padrões da família MPEG que possuem algoritmos de compressão. Autores adotam esses padrões para facilitar a extração de informação em conteúdo multimídia. Benefícios como redução da complexidade

computacional, tempo de processamento e quantidade de memória utilizada são mencionados como fatores que levam a adoção dessa abordagem (NGO; PONG; CHIN, 1998; NGO; ZHANG; PONG, 2001; CALIC; IZQUIERDO, 2002).

Algoritmos que seguem este padrão empregam uma combinação de técnicas de compressão com perda juntamente com compressão sem perda no processo de codificação de vídeo. Ambas as compressões são divididas em três etapas, sendo a etapa de compressão com perda composta pelas etapas de Estimativa de Movimento, Codificação por DCT e Quantização, e as etapas sem perdas composta por Vetorização, Codificação por Entropia e *BitsTream*. Todos os estágios de compressão trazem como vantagem a obtenção de um arquivo comprimido de menor dimensão, mantendo, no entanto, uma qualidade mínima em relação ao original, conforme o objetivo que se pretende.

Reduções temporais e espaciais compõem as técnicas de compressão de dados do padrão. Na compressão temporal utilizam-se diferentes modos de compressão dos dados para para possibilitar a geração do fluxo elementar do vídeo (STANDARDISATION, 2000). Já na compressão espacial são eliminadas informações redundantes buscando identificar as informações repetidas presentes em quadros próximos para codificar apenas um desses quadros, eliminando a codificação da informação nos demais, diminuindo a quantidade de informação a ser codificada no *Bitstream* final (RICHARDSON, 2003).

## 2.6 Considerações Finais

Este capítulo apresentou os conceitos e tecnologias relacionadas ao processo de análise de conteúdo de vídeo digital. A estrutura do vídeo, assim como seus segmentos, também foram descritos de modo que fique mais fácil a leitura dos temas relacionado a essa área e que serão abordados nos próximos capítulos.

Observou-se que a recuperação e manipulação de informação em conteúdo multimídia não é uma tarefa simples. É necessário realizar a extração de dados como cor, textura, movimentação, etc, para segmentar o vídeo em pedaços menores a fim de facilitar seu acesso posterior. Segmentar o conteúdo em cenas, ou unidades semânticas, para identificar conteúdo de alto nível é ainda mais difícil, visto que há uma lacuna entre os dados do vídeo e a interpretação do usuário.

Para avaliação de técnicas que realizam segmentação temporal em vídeos, frequentemente são utilizadas medidas quantitativas como o *precision* e o *recall*, as quais tem o objetivo de verificar a eficiência da metodologia empregada. Entretanto verificou-se que tais medidas possuem limitações, sendo apresentadas outras maneiras de avaliação como derivações dessa medida e técnicas centradas no usuário. Ainda, o uso de padrões de compressão de vídeo são amplamente utilizados para extração de informação em vídeo, sendo apresentado o processo de compressão mais adotado pela comunidade, o padrão MPEG.

Capítulo

3

# Revisão Sistemática

## 3.1 Considerações Iniciais

Esse capítulo tem como propósito descrever e realizar uma pesquisa que forneça apoio e melhore a busca pelo estado da arte no tema apresentado. Para tal faz-se o uso da Revisão Sistemática, definida por Biolchini et al. (2005) como uma metodologia específica de pesquisa, desenvolvida para unir e avaliar as evidências disponíveis a respeito de um determinado tópico.

Corroborando as palavras de Biolchini, Kitchenham (2004) afirma que a revisão sistemática consiste de meios para identificar, avaliar e interpretar toda pesquisa relevante disponível para um determinado problema de pesquisa, tópico de uma área ou fenômeno de interesse. Fica evidente que ambas as definições apresentadas focam em realizar uma pesquisa fazendo análises e tentando extrair informações relevantes sobre o assunto da mesma.

A revisão tradicional difere em muitos pontos da revisão sistemática. A primeira geralmente é escrita por especialistas, onde os métodos de coleta e interpretação dos estudos são informais e subjetivos, criando uma tendência a citar seletivamente literaturas que reforçam noções preconcebidas, além de não possuir uma descrição da pesquisa, seleção e avaliação da qualidade dos estudos (PAI et al., 2004). Na revisão sistemática, a busca é abrangente e exaustiva por estudos primários seguindo uma questão. Critérios de qualificação são reproduzíveis e claros para a seleção de estudos, possuindo uma avaliação explícita, assim como o método, o qual é pré-determinado (KITCHENHAM, 2004)

O objetivo de tal revisão é produzir uma síntese completa de trabalhos publicados sobre uma questão de pesquisa específica, utilizando um processo metodológico bem definido para guiar o procedimento de busca e análise de trabalhos (BIOLCHINI et al., 2005).

Basicamente, a revisão sistemática inclui as tarefas de formulação, extração e análise dos resultados, representadas, respectivamente por:

1. Planejamento da revisão. É identificada a necessidade de tal revisão e, posteriormente, desenvolve-se um protocolo de revisão. Divide-se em duas etapas:
   - Formulação da Questão
   - Seleção de Fontes
2. Condução da Revisão. Com as fontes definidas, a busca é realizada e os estudos obtidos são avaliados de acordo com os critérios criados. Divide-se em duas etapas;
   - Seleção de estudos
   - Extração de informações
3. Análise dos Resultados. Ocorre a sumarização e análise dos resultados, geralmente por meio de métodos estatísticos utilizando componentes como tabelas e/ou gráficos.

A seguir, está documentada uma revisão que segue processos sistemáticos, a qual aborda o tema de manipulação de vídeos digitais, notadamente a identificação de estruturas em conteúdo multimídia denominadas cenas. Nas subseções seguintes são apresentadas as etapas, de modo detalhado, da revisão sistemática seguindo o modelo apresentado por Kitchenham (2004).

## 3.2 Planejamento da Revisão

### 3.2.1 Formulação da Questão

#### Foco da questão

Essa revisão sistemática tem como foco obter técnicas e modelos atuais relacionados à área de segmentação de vídeo, especificamente à recuperação de informação em arquivos de vídeos, utilizando para tal uma semântica acoplada a estes conteúdos a fim de identificar objetos conhecidos como cenas.

#### Qualidade e Amplitude da Questão

- Problema: A manipulação de arquivos multimídias como os vídeos tem sido um desafio na literatura, pois arquivos desse tipo possuem muitas informações associada a seu conteúdo. Atualmente existem técnicas para segmentação de vídeo de modo automático e semi-automática que visam facilitar a extração de informações e seus conteúdos, no entanto, para melhorar tal segmentação têm-se buscado adicionar

semântica a estas técnicas, possibilitando extração de informações mais eficiente e possibilitando identificação de segmentos denominados cenas.

- Questão: Quais desafios das técnicas/modelos/métodos de segmentação de vídeo que faz uso de semântica?
- Palavras-chave e sinônimos: *scene segmentation*, *video segmentation*, *temporal video segmentation*, *temporal segmentation*, *semantic segmentation*, *semantic video segmentation*, *content-based image retrieval*, *content-based video retrieval*, *automatic indexing of video data*, *temporal video segmentation*, *semantic gap*, *MPEG-4*.
- Intervenção: Abordagens (técnicas/métodos/modelos) que realizam segmentação de vídeo, utilizando semântica.
- Efeito: Obtenção do estado da arte e classificação das abordagens para segmentação de vídeo utilizando semântica.
- Métrica do Resultado: Número de estudos identificados.

### 3.2.2 Seleção de Fontes

#### Critérios para Seleção de Fontes

Disponibilidade de artigos na Web; e consideração de fontes relevantes utilizando palavras-chaves.

#### Idioma

Inglês.

#### Identificação de Fontes

- Busca: Buscas em site que oferece serviços de buscas por trabalhos acadêmicos na web (Google Scholar[1], IEEE [2], ACM [3], Elsevier [4], Springer [5]).
- String:

  1) Inicialmente, foram efetuadas buscas de acordo com o escopo do projeto, ou seja, segmentação semântica de vídeo ("video semantic segmentation"). No entanto, não houve um número de resultados satisfatórios.
  *String: Video Semantic Segmentation.*

[1] http://scholar.google.com.br
[2] http://www.ieee.org e http://ieeexplore.ieee.org
[3] Digital Library: portal.acm.org/dl.cfm
[4] www.elsevier.com
[5] www.springer.com

2) Assim, foi necessário generalizar a string para algo como "video segmentation" a fim de entender o processo de segmentação de vídeo sem semântica.
*String: (video segmentation OR (temporal segmentation OR temporal video segmentation)) AND MPEG 4.*
3) Em seguida, foi identificada uma área que utilizava segmentação de vídeo, a CBIR (Content-Based Information Retrieval), a qual realiza a recuperação de informação baseada em conteúdo em arquivos de imagens. Essa área levou ao principal desafio do tema do projeto, a lacuna semântica (semantic gap).
*String: (semantic segmentation OR semantic vídeo segmentation) AND semantic gap AND content based information retrieval.*
4) Finalmente, com estudos mais aprofundados, foi obtida a string final de busca.
*String: (video semantic "shot clustering" OR "scene segmentation" -survey) ou (video AND semantic AND ("shot clustering" OR "scene segmentation" -survey)) .*

- Lista: Foram incluídas fontes de anais de conferências e periódicos publicados pela IEEE, ACM, Springer e Elsevier.

## 3.3 Condução da Revisão

A necessidade de produzir a Condução da Revisão está em identificar os estudos relacionados ao tema da pesquisa, utilizando uma estratégia de pesquisa abrangente. Os tópicos abordados nesta fase exigem em certo rigor para que o processo de pesquisa tenha um resultado satisfatório.

### 3.3.1 Seleção de Estudos

**Definição de Estudos**

- Critérios de inclusão e exclusão: os critérios adotados para a seleção dos trabalhos fazem referência às técnicas/abordagens utilizados no tema do trabalho em questão. Não criou-se critério de exclusão referente a data, incluindo, portanto, todos os trabalhos existentes na literatura. A exclusão de *surveys* também foi efetuada, pois o intuito da revisão é realizar uma pesquisa mais abrangente e exaustiva, o que algumas vezes não ocorre com este tipo de trabalho.

- Definição de tipos de estudo: Todo e qualquer trabalho/artigo relacionado ao tópico, que siga os critérios de inclusão e exclusão, serão selecionados.

- Procedimento para seleção de estudos: A string de busca foi inserida no sítio Google Acadêmico (Google Scholar) utilizando sua ferramenta de busca avançada. A Figura 3.1 ilustra a pesquisa sendo realizada no sítio Google Acadêmico. Esse sítio tem

como característica incluir a base de dados dos principais periódicos acadêmicos existentes. Para validar o processo de pesquisa, a string também foi inserida nos mecanismos de buscas dos principais periódicos relacionados a área de computação e a análise de vídeos (ACM, IEEE, Elsevier e Springer) a fim de comparar os resultados com os obtidos com o textitGoogle Scholar. Os trabalhos selecionados foram armazenados em um programa de código aberto (do inglês, *Open Source*) denominado JabRef [6], possibilitando a extração e importação de referências em formato de arquivo .bib, o qual facilita o processo de citação e referências no editor de texto LaTeX .Em seguida, é efetuada uma pré-seleção dos trabalhos que tem como critério a análise dos títulos dos trabalhos. Após, seleção é realizada pela leitura do resumo (do inglês, *abstract*) e, se necessário, da conclusão e, por fim, e feito leitura total dos artigos para extrair dados quantitativos ao tema de pesquisa.

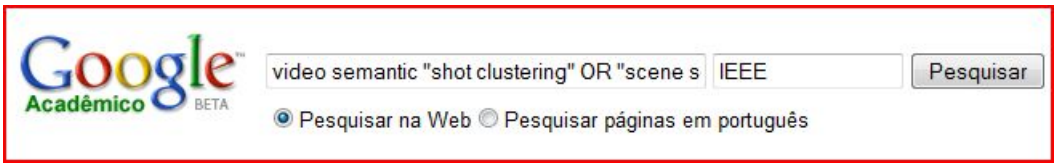



Figura 3.1: Exemplo de busca no Google Acadêmico

### 3.3.2 Execução da Seleção

#### Seleção de Estudos Iniciais

Para a fase de seleção dos estudos iniciais foram efetuadas buscas, nos sítios já mencionados, com a seguinte string: (video semantic “shot clustering” OR “scene segmentation” -survey) ou (video AND semantic AND (“shot clustering” OR “scene segmentation” -survey)). Contudo, devido aos problemas encontrados na composição da string de busca nos periódicos da IEE e ACM, os resultados desses dois foram descartados da análise, já com os sítios da Elsevier e Springer não houve dificuldades na construção da string, todavia, a quantidade de resultados retornados foi inferior ao resultado obtido pelo Google Acadêmico. Em seu mecanismo de busca, o Google leva em conta o texto integral de cada artigo, o autor, a publicação em que o artigo saiu e a frequência com que foi citado em outras publicações acadêmicas. Assim, o sítio do Google Acadêmico foi definido como o único meio de pesquisa dos trabalhos, tanto pela facilidade da construção da string, quanto pela maior quantidade de trabalhos obtidos. A seguir, são apresentados os resultados do processo de busca, assim como os problemas encontrados para inserir a string definida e algumas alternativas (compondo novas strings) para tentar resolvê-los.

Os resultados obtidos do processo de inserção de string são apresentados a seguir:

[6] http://jabref.sourceforge.net/

1. Google Scholar (12/09/2008)

   String: (video semantic "shot clustering" OR "scene segmentation" -survey)

   - IEEE: 149 trabalhos selecionados.
   - ACM: 39 trabalhos selecionados.
   - Springer: 93 trabalhos selecionados.
   - Elsevier: 27 trabalhos selecionados.

2. IEEE (13/09/2008)

   Não foi possível incluir mais de uma frase para a operação de OU, além do mecanismo "should contain" não possuir relevância alguma no resultado da pesquisa.

   String: "scene segmentation", "shot clustering", +video, +semantic, +scene, -survey

   - 224 trabalhos encontrados

   String (IEEExplore): ((video <and >semantic <and >( shot clustering <or >scene segmentation ) <not >survey) <in >metadata)

   - 18 trabalhos encontrados

   Para refinar o resultado, as palavras semantic e scene foram pesquisadas apenas nas descrições dos arquivos. String: "scene segmentation" , "shot clustering", +description:semantic, +description:scene, -survey

   - 90 trabalhos encontrados

   A palavra vídeo foi adicionada na descrição da string. String: "scene segmentation", "shot clustering", +description:semantic, +description:scene, +description:video, -survey

   - 42 trabalhos encontrados

3. ACM - DIGITAL LIBRARY (15/09/2008)

   2 strings de busca foram utilizadas devido a falta do operador lógico "OU" para realizar busca com frases contendo as expressões shot clustering e scene segmentation.

   String: (semantic, and video) and ("shot clustering") and (not survey)

   - 17 trabalhos encontrados

   String: (semantic, and video) and ("scene segmentation") and (not survey)

   - 25 trabalhos encontrados

4. Springer (15/09/2008)

   O mecanismo de construção da string é semelhante ao Google Acadêmico, facilitando a pesquisa.

   String: video and semantic and ("shot clustering" or "scene segmentation") and not survey

   - 74 trabalhos encontrados

5. Elsevier (16/09/2008)

   String: video AND scene AND semantic OR "shot clustering" OR "scene segmentation" AND NOT survey

   - 23 trabalhos encontrados

### 3.3.3 Extração de Informações

#### Critérios de inclusão e exclusão de informações

As informações obtidas devem mencionar a técnicas/abordagens para segmentação ou identificação de cenas em vídeos digitais, podendo utilizar de segmentos de vídeos denominados tomadas. Entretanto, esses segmentos (tomadas) não podem ser o foco da pesquisa, mas pode ocorrer de serem citados em alguns trabalhos. Assim, o estudo desses segmentos (identificação, mudança, segmentação) é deixado em segundo plano no processo de extração de cenas.

#### Formulários de Extração de Dados

O formulário de extração de dados tem como objetivo auxiliar na obtenção dos resultados desse estudo. Para isso, dados referentes às técnicas e algoritmos adotados devem ser extraídos dos trabalhos selecionados. Dentre esses dados estão:

- Título.
- Gênero(s) dos vídeos avaliados.
- Mídias que foram exploradas (visual, áudio, texto).
- Tecnologia explorada (e.g. MPEG, Flash).
- Ano de publicação.

**Execução de Extração**

Os resultados obtidos do estudo e análise das abordagens estudadas serão apresentados na próxima seção.

## 3.4 Resultados

O objetivo deste trabalho está em identificar e analisar técnicas que reconheçam e identifiquem estruturas de vídeo denominadas cenas. Tais estruturas necessitam que seja determinada algum tipo de semântica acoplada ao conteúdo do arquivo de vídeo. Portanto, serão selecionados trabalhos que possuem maior relevância na área de recuperação de informação em vídeo, especificamente os que tentam extrair cenas utilizando alguma semântica em sua técnica/algoritmo.

### 3.4.1 Seleção dos trabalhos

Foram selecionados, com as strings de busca formalizadas durante as etapas de planejamento e condução da revisão, 297 trabalhos da área relacionada ao tema. Na Figura 3.2 pode-se observar a quantidade de trabalhos obtidos considerando as fontes definida na pesquisa. Nessa figura fica evidente que a IEEE possui grande interesse nessa área de recuperação de informação audiovisual, visto que ele possui quase metade dos trabalhos analisados.

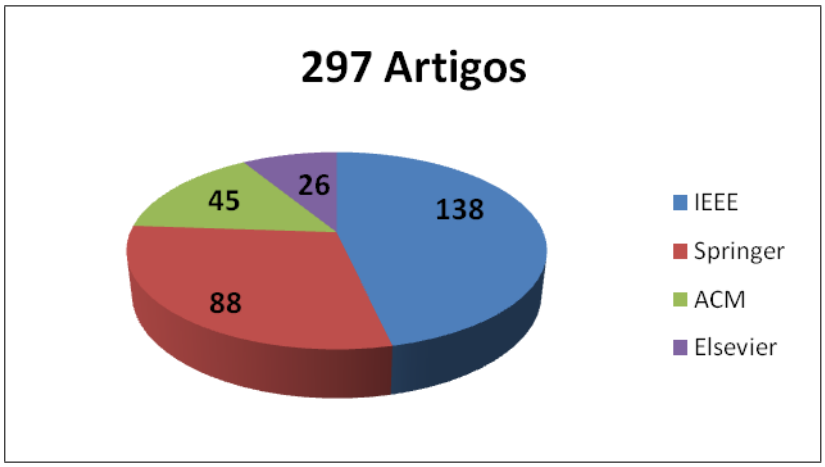



Figura 3.2: Distribuição do número de artigos por periódico

É possível observar na Figura 3.3 que mais da metade da produção de conteúdo científico nessa área ocorreu durante os últimos 4 anos, deixando evidente a relevância do tema na atualidade.

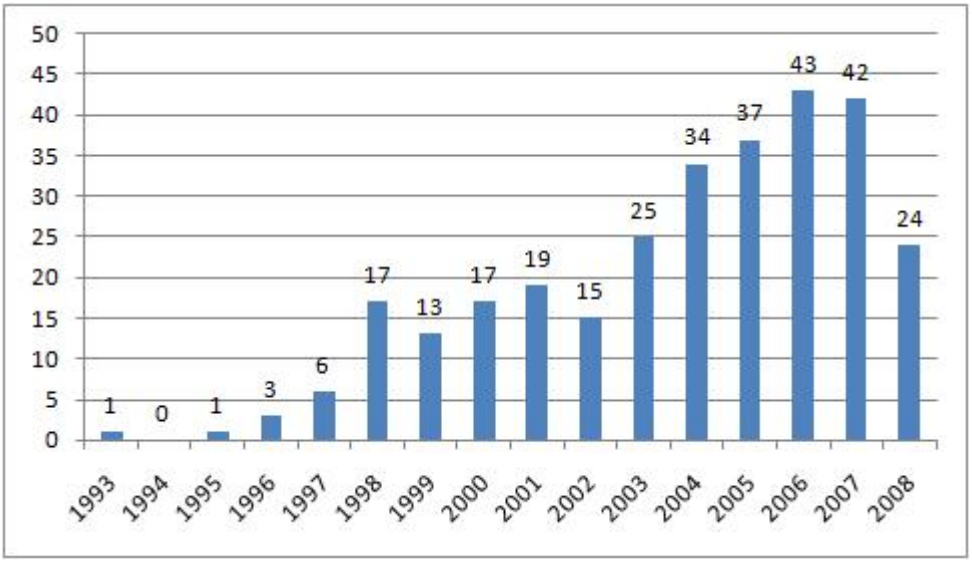



Figura 3.3: Distribuição da quantidade de artigos pelo tempo

A Figura 3.4 ilustra a produção de todas as fontes. Entretanto, observa-se que IEEE e Springer são os maiores contribuintes na contribuição da produção científica de trabalhos na área ao longo do tempo.

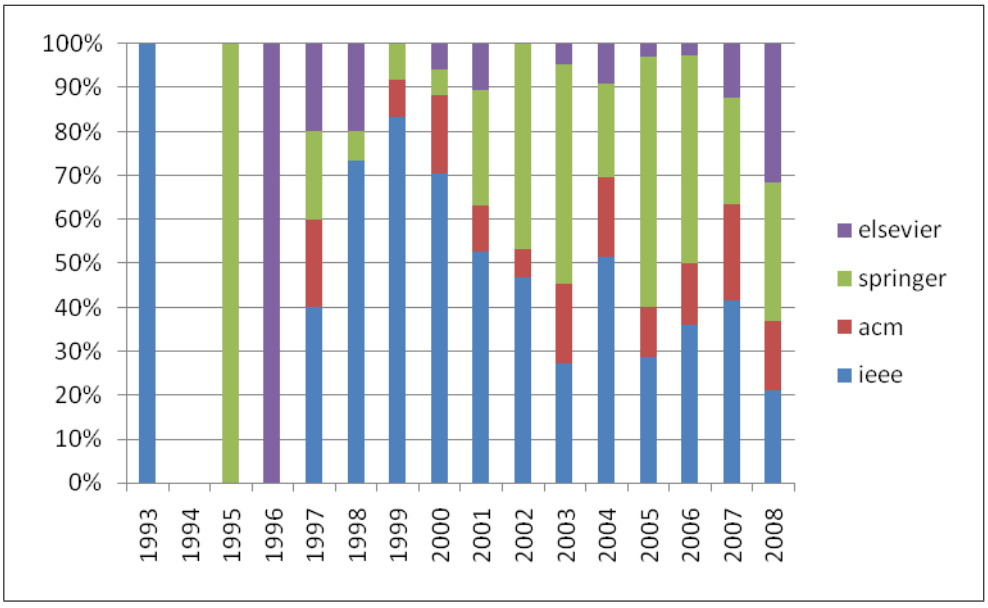



Figura 3.4: Porcentagem de produção no decorrer do tempo

Afim de refinar o número de trabalhos, deixando apenas os mais relevantes, foi utilizado o critério de exclusão por título, o qual consistia na leitura do título do artigo para certificar sua relevância com o tema. Nessa fase, foram excluídos 70 trabalhos, restando um total de 227. Para excluir os trabalhos, algumas palavras ou sentenças foram consideradas: spatial segmentation, shot detection, region segmentation, object segmentation, shot boundary detection, face recognition, syntactical segmentation, video navigation. Com esse resultado é possível inferir que a *string* de busca foi bem definida, visto que retornou

poucos trabalhos irrelevantes.

Na etapa final, foram selecionados trabalhos que continham em seus resumos palavras como: scene, semantics, high level information, clustering, video retrieval, video events, video summarization, video analysis. Assim, do conjunto de 227 trabalhos, 88 foram selecionados.

A Figura 3.5 apresenta a quantidade de trabalhos por periódico foi obtido após a fase de seleção por resumo.

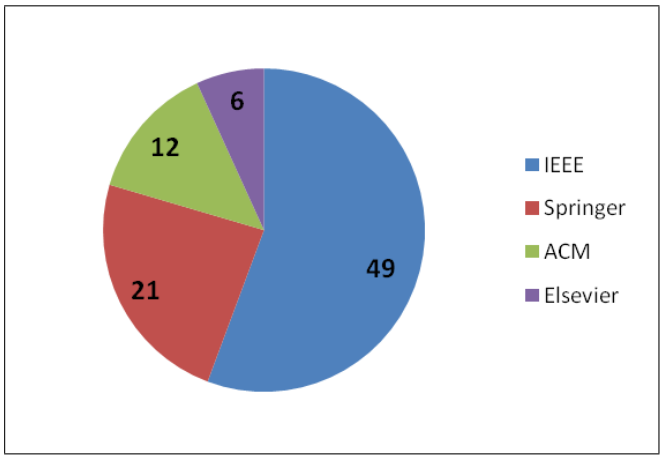



Figura 3.5: Distribuição do número de artigos por periódico, após fase de seleção por resumo

Continuando a demonstração dos resultados quantitativos, a Figura 3.6 evidencia que a pesquisa é realizada em cima de um tema muito procurado pela comunidade científica dos últimos 5 anos atrás até esta presente data (09/2008).

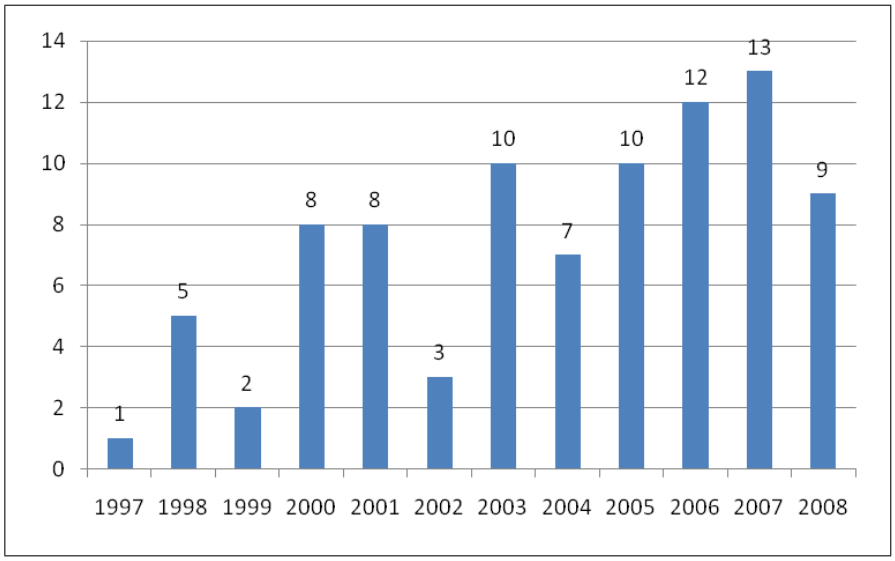



Figura 3.6: Distribuição da quantidade de artigos pelo tempo, após fase de seleção por resumo

Outros dados, além de datas e fontes, podem ser extraídos dos trabalhos selecionados a fim de definir os conceitos e técnicas importantes na análise. Gênero do vídeo avaliado,

mídia(s) utilizadas na técnica e tecnologia empregada (se existir) são exemplos de dados que podem contribuir com a análise na obtenção dos resultados. A Figura 3.7 apresenta a quantidade e a descrição de todos os gêneros de vídeos (16 no total) avaliados nos trabalhos selecionados. Como muitos desses trabalhos validavam a técnica em nenhum ou mais de um gênero, todos foram computados, de maneira que se aparecer um trabalho com três gêneros, cada um representa uma unidade no gráfico da 3.6. O interessante nessa figura é a pouca atenção que a área social proporciona (educação, saúde e segurança), além dos trabalhos exaustivos realizados em gêneros como filme, noticiário e esporte.

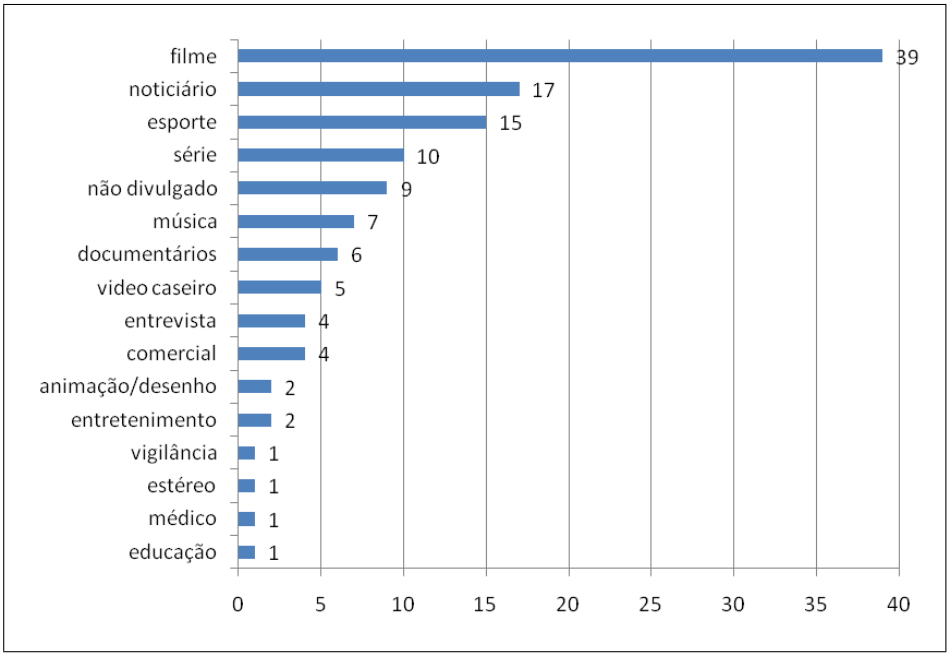



Figura 3.7: Gêneros de vídeos identificados no processo de validação das técnicas

Outra característica importante são as mídias usadas pelas técnicas, cada autor adota uma metodologia de acordo com as mídias que pretende processar para extrair informação. A Figura 3.8 lustra que, apesar de ser o mais antigo método de extração de informação em vídeo, as características visuais ainda predominam quando o assunto é segmentação de cena. Merecem destaques também as técnicas que utilizam mais de um tipo de mídia, a qual teoricamente, tem um resultado melhor, mas torna-se mais complexo.

Em cerca de quase 40% dos trabalhos analisados a metodologia adotada foi a adoção de um formato de compressão, especificamente o MPEG. Outro formato de mídia extraído foi o Flash$^{TM}$ [7], o qual é um formato de mídia baseado em vetores mas que fornece macanismos que geram imagens em movimento (3 %). Entretanto, os vídeos que não possuem compressão de dados são amplamente utilizados (Figura 3.9).

Todos os dados e análises visualizados anteriormente foram obtidos por intermédio do formulário de extração de dados descritos na seção anterior sobre a condução da revisão.

[7] http://www.macromedia.com/software/flash/about/

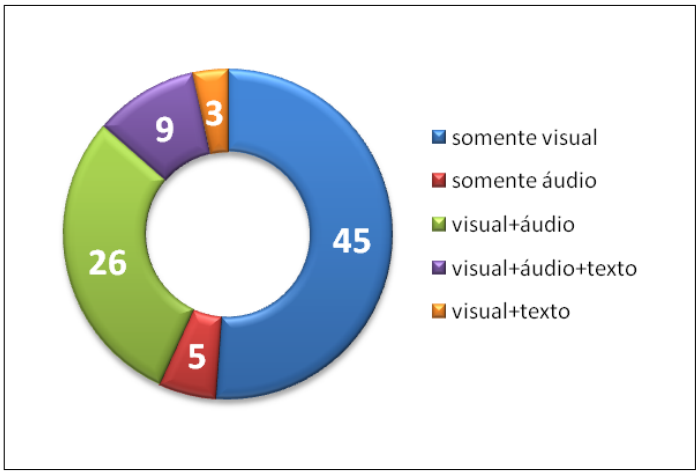



Figura 3.8: Quais tipos de mídias foram extraídas em cada trabalho

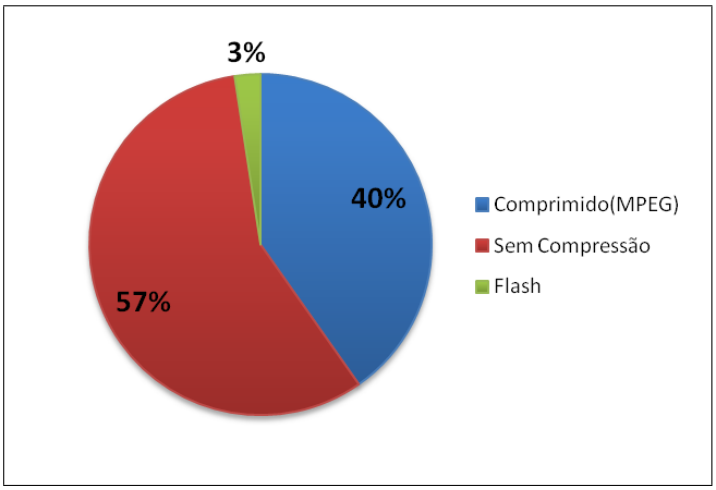



Figura 3.9: Quantidade dos trabalhos que utilizam um formato de compressão(em %)

O formulário de extração de dados desenvolvido fica evidente na Tabela 3.1, a qual sumariza os oitenta e oito (88) artigos selecionados. Na coluna correspondente à codificação, o símbolo *U* significa que a metodologia faz uso de vídeo sem compressão (do inglês, *Uncompressed*).

Tabela 3.1: Tabela de extração de dados dos artigos selecionados

| **Título** | **Gênero(s)** | **Mídia(s)** | **Codif.** | **Ano** |
|---|---|---|---|---|
| A cinematic-based framework for scene boundary detection in video | filme, documentário | visual | MPEG | 2003 |
| A computational approach to semantic event detection | documentário | visual | MPEG | 1999 |
| A framework for segmentation of talk and game shows | entrevistas, comerciais | visual | *U* | 2001 |
| A hierarchical approach to scene segmentation | série, filme | visual | *U* | 2001 |
| A method for scene structure construction with time constraint | filme | visual | *U* | 2003 |

| | | | | |
|---|---|---|---|---|
| A Method of Generating Table of Contents for Educational Videos | educacional | visual+áudio+ texto | MPEG-7 | 2005 |
| A motion Based Scene Tree for browsing and retrieval of compressed videos | não informado | visual | MPEG-2 | 2004 |
| A Multi-expert System for Movie Segmentation | filme | visual+áudio | MPEG | 2002 |
| A Multi-Modal Approach to Story Segmentation for News Video | noticiário | visual+áudio | MPEG | 2003 |
| A Multimodal Scheme for Program Segmentation and Representation in Broadcast Video Streams | não informado | visual+áudio+ texto | MPEG | 2008 |
| A new approach for high level video structuring | filme | visual | MPEG-1 | 2000 |
| A novel scheme for video scenes segmentation and semantic representation | entrevista, filme | visual | *U* | 2008 |
| A semantic model for flash retrieval using co-occurrence analysis | filme, música | visual+texto | Flash | 2003 |
| A Thousand Words in a Scene | não informado | visual+texto | *U* | 2007 |
| A Two-Layer Graphical Model for Combined Video Shot and Scene Boundary Detection | série | visual+texto | *U* | 2006 |
| A unified framework for semantic shot representation of sports video | esporte | visual+áudio+ texto | *U* | 2005 |
| A visual model approach for parsing colonoscopy videos | médico | visual+áudio | MPEG-2+*U* | 2004 |
| Algorithm of Scene Segmentation Based on SVM for Scenery Documentary | documentário | visual | MPEG | 2007 |
| An Adaptive and Efficient Unsupervised Shot Clustering Algorithm for Sports Video | esporte | visual | *U* | 2008 |
| An Error-Tolerant Video Retrieval Method Based on the Shot Composition Sequence in a Scene | esporte | visual | *U* | 2007 |
| An SVM Framework for Genre-Independent Scene Change Detection | entrevista, noticiário, série, música | visual+áudio | MPEG | 2007 |
| Associating characters with events in films | filme | visual+áudio+ texto | MPEG-7 | 2007 |
| Audio Elements Based Auditory Scene Segmentation | entretenimento | áudio | *U* | 2006 |
| Audio Feature Extraction and Analysis for Scene Segmentation and Classification | comercial, esportes, noticiários | áudio | *U* | 1998 |
| Audio-based description and structuring of videos | filme | áudio | *U* | 2006 |
| Automatic music video generation based on temporal pattern analysis | vídeo caseiro, música | visual+áudio | *U* | 2004 |
| Automatic scene detection for advanced story retrieval | entrevista, filme, animação | visual | MPEG-1 | 2008 |

| | | | | |
|---|---|---|---|---|
| Automatic Scene Detection in News Program by Integrating Visual Feature and Rules | noticiário | visual | *U* | 2001 |
| Automatic summarization of music videos | música | visual+áudio | *U* | 2006 |
| Automatic Video Scene Extraction by Shot Grouping | esporte | visual | MPEG-1 e MPEG-2 | 2000 |
| Automatic video summarizing tool using MPEG-7 descriptors for personal video recorder | música, filme | visual | MPEG-7 | 2003 |
| Computable scenes and structures in films | filme | visual+áudio | MPEG-1 | 2002 |
| Constructing and Application of Multimedia TV News Archives | noticiário | visual+áudio+ texto | *U* | 2007 |
| Content aware video presentation on high-resolution displays | não informado | visual+áudio | MPEG-2 | 2008 |
| Content-based movie analysis and indexing based on audiovisual cues | filme | visual+áudio | MPEG-1 | 2004 |
| Content-based retrieval of video data by the grammar of film | filme | visual | *U* | 1997 |
| Detection of Documentary Scene Changes by Audio-Visual Fusion | documentário | visual+áudio | *U* | 2003 |
| Determining computable scenes in films and their structures using audio-visual memory models | filme | visual+áudio | *U* | 2000 |
| Dissolve transition detection algorithm using spatio-temporal distribution of MPEG macro-block types | noticiário, esporte | visual | MPEG | 2000 |
| Effective Fades and Flashlight Detection Based on Accumulating Histogram Difference | noticiário | visual | MPEG-2 | 2006 |
| Efficient summarization of stereoscopic video sequences | estéreo | visual | *U* | 2000 |
| EMS: Energy Minimization Based Video Scene Segmentation | vídeo caseiro, filme | visual | *U* | 2007 |
| Enhanced Eigen-Audioframes for Audiovisual Scene Change Detection | noticiário | visual+áudio | *U* | 2007 |
| Event detection using multimodal feature analysis | esporte, série | visual+áudio | *U* | 2005 |
| Exploring video structure beyond the shots | filme | visual | MPEG | 1998 |
| Extraction of Film Takes for Cinematic Analysis | filme | visual | *U* | 2005 |
| Finding structure in home videos by probabilistic hierarchical clustering | vídeo caseiro | visual | MPEG-1 | 2003 |
| Fuzzy shot clustering to support networked video databases | música | visual | MPEG | 1998 |
| Hierarchical Summarization of Videos by Tree-Structured Vector Quantization | noticiário | visual | *U* | 2006 |
| Home video structuring with a two-layer shot clustering approach | vídeo caseiro | visual | MPEG e MPEG-7 | 2008 |

| | | | | |
|---|---|---|---|---|
| iJADE surveillant–an intelligent multi-resolution composite neuro-oscillatory agent-based surveillance system | não informado | visual | MPEG-7 | 2003 |
| Investigation on unsupervised clustering algorithms for video shot categorization | esportes | visual | *U* | 2007 |
| Joint scene classification and segmentation based on hidden Markov model | comerciais, noticiário, esporte | visual+áudio | *U* | 2005 |
| Low-level cross-media statistical approach for semantic partitioning of audio-visual content in a home multimedia environment | série, filme | visual+áudio | MPEG-1 + *U* | 2004 |
| Modeling scenes with local descriptors and latent aspects | não informado | visual | *U* | 2005 |
| Movie scene segmentation using background information | vídeo caseiro, filme | visual | MPEG-1 | 2008 |
| Multimedia content analysis-using both audio and visual clues | survey/review | | | 2001 |
| Multimodal Data Fusion for Video Scene Segmentation | filme | visual+áudio+ texto | *U* | 2000 |
| Multi-Modal Dialog Scene Detection Using Hidden Markov Models for Content-Based Multimedia Indexing | filme, série | visual+áudio | MPEG-7 | 2001 |
| NBR:A Content-Based News Video Browsing and Retrieval System | noticiário | visual+áudio+ texto | MPEG | 2007 |
| On fuzzy clustering and content based access to networked video databases | música | visual | MPEG | 1998 |
| On Retrieval of Flash Animations Based on Visual Features | não informado | visual | Flash | 2008 |
| Retrieval of movie scenes by semantic matrix and automatic feature weight update | filme | visual | *U* | 2008 |
| Scene Boundary Detection by Audiovisual Contents Analysis | filme | visual+áudio | *U* | 2005 |
| Scene change detection by audio and video clues | filme, noticiário | visual+áudio | *U* | 2002 |
| Scene Determination Based on Video and Audio Features | filme | visual+áudio | MPEG-1 | 2001 |
| Scene Segmentation and Categorization Using NCuts | filme, série | visual | *U* | 2007 |
| Scene Segmentation and Image Feature Extraction for Video Indexing and Retrieval | não informado | visual | MPEG | 1999 |
| Segmentation, categorization, and identification of commercial clips from TV streams using multimodal analysis | comercial | visual+áudio+ texto | MPEG-1 | 2006 |
| Semantic Segmentation of Documentary Video using Music Breaks | documentário | visual+áudio | *U* | 2006 |

| | | | | |
|---|---|---|---|---|
| Semantic video content abstraction based on multiple cues | filme, série | visual+áudio+ texto | *U* | 2001 |
| Shot clustering techniques for story browsing | filme | visual | MPEG-1 | 2004 |
| Sports video summarization and adaptation for application in mobile communication | esporte | visual+áudio | MPEG-1 | 2006 |
| Structural and Semantic Analysis of video | survey/review | | | 2000 |
| Summarizing Video: Content, Features, and HMM Topologies | série, filme | visual+áudio | *U* | 2003 |
| Temporal Shot Clustering Analysis for Video Concept Detection | esporte | visual | *U* | 2005 |
| The ToCAI Description Scheme for Indexing and Retrieval of Multimedia Documents | esporte | visual+áudio | *U* | 2001 |
| Video browsing and retrieval based on multi-modal integration | esporte, noticiário, filme, comercial | visual+áudio+ texto | *U* | 2003 |
| Video Cataloging System for Real-Time Scene Change Detection of News Video | noticiário | visual | MPEG-1 | 2005 |
| Video Hierarchical Structure Mining | não informado | visual | *U* | 2006 |
| Video scene detection using slide windows method based on temporal constrain shot similarity | não informado | visual | MPEG-7 | 2001 |
| Video Scene Retrieval with Sign Sequence Matching Based on Audio Features | entretenimento, noticiário | áudio | *U* | 2005 |
| Video scene segmentation using Markov chain Monte Carlo | filme | visual | *U* | 2006 |
| Video scene segmentation using video and audio features | filme | visual+áudio | *U* | 2000 |
| Video scene segmentation via continuous video coherence | noticiário | visual | *U* | 1998 |
| Video summaries and cross-referencing through mosaic-based representation | esporte | visual | MPEG | 2004 |
| Video Summarisation for Surveillance and News Domain | vigilância, noticiário | visual | MPEG-7 | 2004 |
| Video summarization by redundancy removing and content ranking | série, filme, desenho | visual+áudio | *U* | 2007 |

## 3.5 Considerações Finais

A construção de estruturas que permitam um crescimento massivo de conteúdo de vídeo mostra-se necessária visto que a demanda por este material cresce a cada dia. Em decorrência do aumento ao acesso desse conteúdo muitos dos problemas associados com a forma de vídeo digital precisam ser investigados. A habilidade de facilmente cap-

turar, formatar, comprimir, armazenar, transmitir e apresentar o vídeo em dispositivos computacionais, sejam eles móveis ou não, faz parte do cotidiano do usuário doméstico. Entretanto, manipular vídeo em formato digital como fonte de informação fornece muitos desafios, além de conhecimento específico e sólido. Entende-se como manipulação de informação os temas de indexação, análise, sumarização, agregação, navegação e pesquisa.

Ficou evidente que no contexto do vídeo digital muitas dessas tarefas dependem do domínio de onde o vídeo é gerado. Por exemplo, a análise que é feita num evento esportivo é diferente da análise de um programa jornalístico. Desse modo, diferentes metodologias de extração de informação são utilizadas em diferentes formatos de vídeos. Ainda, uma das maneiras relatadas nessa revisão foi a análise semântica do conteúdo em estruturas temporais do tipo cena. Tais estruturas são formadas de partes denominadas tomadas. Enquanto as tomadas são partes sintéticas, as cenas representam em nível semântico de determinadas partes do vídeo.

Neste levantamento observou-se que há inúmeras técnicas/algoritmos/ferramentas que empregam o uso de identificação/segmentação de cena para análise, recuperação e indexação de vídeo. O estudo desse tipo de estrutura é essencial para que a manipulação do conteúdo seja feita de modo eficiente e correto, assim como a presença de padrões de compressão (MPEG) no conteúdo multimídia.

Segundo os dados coletados e analisados por meio desta revisão pode-se observar que o assunto de cena em vídeos digitais é um estudo válido visto que está ocorrendo uma grande produção científica nos últimos anos até o momento, demonstrando que o tema tem uma importância considerável no desenvolvimento de aplicações multimídia, em especial quando os vídeos digitais estão inseridos. A identificação de cenas está acontecendo recentemente por meio de técnicas multimodais, ou seja, em técnicas que abordam mais de um tipo de mídia (imagem, texto, áudio), tendo os resultados mais expressivos estão relacionados a este tipo de técnica.

Como resultado obteve-se um panorama geral de qual tipo de dados que as pesquisas realizadas empregam, sejam eles relacionados à codificação, ao gênero ou às mídias que participam dessa análise.

Capítulo

4

# Trabalhos Relacionados

## 4.1 Considerações Iniciais

Inúmeras técnicas e algoritmos vêm sendo desenvolvidos para facilitar o acesso e recuperação de conteúdo multimídia ao longo das últimas décadas. Áreas como visualização computacional, processamento digital de imagens, reconhecimento de padrões, hipermídia [1], processamento de linguagem natural, entre outras, demandam grandes esforços para diminuir a lacuna entre a capacidade de interpretação humana e o a capacidade de interpretação que os computadores realizam nos dados dos vídeos.

Uma quantidade considerável de características fazem parte das pesquisas que tentam preencher essa lacuna, sendo encontrados em temas relacionados a identificação de cenas ou segmentos de alto nível. Essas características permitem diversas classificações e são representadas pelos gêneros de vídeos (esporte, filme, noticiário, documentário, etc.) ou ainda, caso exista, o tipo de tecnologia abordada (e.g. padrões MPEG). Contudo, a classificação utilizada nesse capítulo foi de acordo com as mídias adotadas nos trabalhos, isto é, imagens, áudio, texto ou a combinação de mais de uma delas. A seguir, serão descritos esses trabalhos usando a classificação mencionada.

## 4.2 Características Visuais

Dentre as áreas responsáveis pela recuperação de conteúdo, a Recuperação de Conteúdo Baseada em Imagem (do inglês, *Content Based Image Retrieval*) é consider-

[1] Hipermídia é a área responsável por desenvolver aplicações que usam relacionamentos associativos entre informações contidas em dados de múltiplas mídias, visando facilitar acesso e manipulação de informações encapsuladas nos dados (LOWE; HALL, 1999).

ada o gargalo na recuperação de conteúdo multimídia (DEB; ZHANG, 2004). Isso porque a maior dificuldade está em interpretar o conteúdo das imagens, pois cada pessoa pode interpretá-las de maneira distintas, ocasionando uma subjetividade que torna difícil o trabalho de extração de informação realizado pelo computador.

Um elemento importante na extração de semântica utilizando características visuais é o quadro-chave (do inglês, *key-frame*). Geralmente, eles são criados após a segmentação de tomadas, pois cada quadro-chave representa uma tomada. Portanto, as técnicas que criam esses quadros devem ser eficientes, uma vez que originam deles a estrutura da próxima camada, isto é, a cena. Zhu e Liu (2008b) detectam e removem os quadros-chaves que não possuem informação útil, criando cenas a partir de similaridade visual e temporal das tomadas que não tiveram seus quadros excluídos. Cao (2007) extrai a cor e a textura dos quadros-chaves para classificadores binários baseados em SVM (do inglês, *Support Vector Machine*) para separar o vídeo (MPEG) em diferentes classes semânticas.

Segundo Smeaton (2007), quanto maior o número de características de baixo nível extraídas de uma imagem, maior a possibilidade de extração semântica de informação. Baseado nessa afirmação algumas pesquisas estão adotando esse método para identificar cenas em vídeos. Yoo (2008) extraiu vinte características de baixo nível de um vídeo, as quais são empregadas na construção de uma matriz de relevância por meio de algoritmos genéticos. Usuários participam de interações nessa matriz para determinar um nível de eficiência maior na recuperação de informação semântica. Chen, Lai e Liao (2008) também adotam o método de extração de muitas características na determinação de similaridade de tomadas, a qual compõe, juntamente com o processo de agrupamento de tomadas, as duas etapas principais da extração de cena. Nessa abordagem, ocorre o princípio de que tomadas possuem imagens de fundo similares, possibilitando a construção da imagem de fundo de uma cena com a técnica do mosaico.

Ainda, há metodologias que analisam as similaridades das tomadas adjacentes para definir o grupo relacionado a uma determinada cena. Zhang e Jiang (2008) adotam uma abordagem de duas camadas levando em conta as similaridades das tomadas adjacentes para solucionar problemas relativos a extração de características e organização do grupo de tomadas. Já Gu et al. (2007) levam em consideração as distribuições locais e globais da continuidade temporal para minimizar o custo computacional, assim como as similaridades das tomadas adjacentes.

## 4.3 Características de Áudio

O uso mais comum do áudio para determinar a semântica do vídeo é servir como um auxílio aos métodos que já utilizam outra mídia. A detecção de silêncio, por exemplo, é utilizada em conjunto com técnicas de recuperação baseada em imagens e texto. Em gêneros como filmes, a escolha do áudio torna-se apropriada visto que são geradas muitas

ambiguidades visuais na transição de segmentos semânticos com esse gênero (HANJALIC et al., 2001).

Mas mesmo em conteúdo multimídia como o vídeo a extração de informação semântica somente com áudio é possível, apesar da forma de fazê-lo não ser tão intuitiva se comparada às mídias visual e texto. Uma possível explicação para essa afirmação pode ser a tendência das pessoas focarem mais atenção nas imagens em movimentos que no próprio som, que fica em segundo plano. De acordo com Harb e Chen (2006) é possível detectar mudanças de cenas em filmes mesmo sem acesso ao conteúdo visual do vídeo, apenas com o som. Ainda, esse mesmo autor diz que é possível estruturar o vídeo mesmo que a fala (linguagem) não seja a mesma da pessoa que está escutando.

Exemplos em que somente o fluxo de áudio pode ocasionar mudanças de significado de alto nível podem ocorrer na trilha sonora de um diálogo, que representa a fala de pessoas, seguida por uma mudança de som no plano de fundo ou ainda mudança nos tópicos dos noticiários de televisão ocasionado pela chamada da notícia pela apresentador do telejornal. Jiang, Lin e Zhang (2000) relata que é possível realizar detecção de transição de cenas por intermédio de segmentos elementares da trilha de áudio, os quais são classificados em segmentos de fala (voz de pessoas) e segmentos sem fala (música, som ambiente e silêncio)(Figura 4.1).

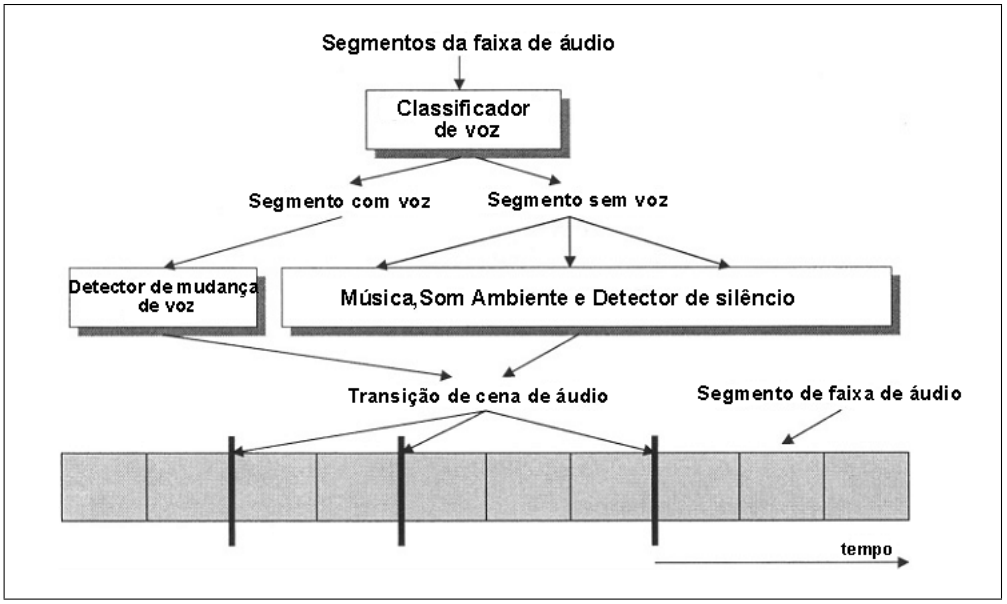



Figura 4.1: Visão geral dos métodos de detecção de cena com áudio por Jiang, Lin e Zhang (2000)

Pode-se dizer que os trabalhos mais recentes envolvendo detecção de cenas somente com áudio realizam o processo de extração de semântica quando conseguem classificar o som em eventos coerentes com o conteúdo (gênero) do vídeo avaliado. Lu, Cai e Hanjalic (2006) adota uma abordagem análoga a recuperação de texto por palavras-chave, classificando os elementos de áudio (música, som ambiente e voz) em onze categorias: fala de três

pessoas diferentes, músicas, barulho, música com fala, aplauso com fala e aplausos com três músicas diferentes. Essa abordagem auxilia na detecção das transições de cenas, no entanto, a relação de transição de cenas existentes com a presença de transições que não eram claras ficou na proporção de três para duas, respectivamente, reduzindo a eficácia do método.

A classificação do áudio em segmentos de vozes e som de fundo é a etapa inicial do método apresentado por Morisawa, Nitta e Babaguchi (2005). Um vetor de características armazena esses segmentos em grupos, os quais são transformados em índices. As cenas com seus respectivos índices são comparadas com amostras de som do vídeo original. A avaliação do resultado foi feita em dois gêneros, entretenimento e noticiários. No entretenimento, a comparação conseguiu identificar segmentos com fala, mas não a fonte da voz corretamente. Em ambos, notou-se que houve sucesso em detectar cenas com música, muito provavelmente devido aos segmentos correspondentes ao som de fundo. Uma vantagem nesse método é a velocidade da recuperação de cenas, menos de 1 segundo por cada sequência comparada.

Harb e Chen (2006) classificam o conteúdo sonoro em seis categorias: diálogo, diálogo calmo, emoção, medo, ação natural e efeitos especiais. Cada unidade sonora (tomada e/ou cenas) é identificada por uma combinação de informações de áudio e adicionada às categorias. Foram avaliados quatro filmes, obtendo uma eficiência de quase 80% em sua metodologia. Mesmo que os resultados sejam estimados para segmentos que possuam por volta de 60 segundos, segmentos com tamanhos maiores ocasionam poucos erros. Entretanto, em gêneros como filmes esses segmentos maiores aparecem constantemente.

## 4.4 Características Audiovisuais

Como visto no capítulo anterior, a maior parte das pesquisas para segmentação de cenas é relacionada ao uso de características visuais. Entretanto, é possível inferir semântica mais confiável usando outros dados do vídeo, como o áudio, por exemplo (HARB; CHEN, 2006). Além da complexidade do processamento do áudio ser menor, é possível salvar processamento das características visuais usando o áudio para definir algumas respostas definitivas considerando o conteúdo da cena (WANG; LIU; HUANG, 2000).

Com o benefício de incluir o som na recuperação de arquivos multimídias, Shao et al. (2006) adotou uma forma de sumarizar vídeos musicais. O método consiste em separar a trilha de música da trilha de vídeo, aplicar técnicas para sumarização na música e detecção de tomadas no vídeo e alinhá-los posteriormente. A avaliação foi centrada no usuário e obteve resultados comparáveis a sumarização manual. Já a pesquisa de Dong e Li (2006) faz uso das mesmas técnicas no fluxo de áudio, como o *zero-crossing rate*, que os autores anteriores e visa detectar segmentos semânticos em documentários por meio de intervalos sonoros. Um desafio nesse trabalho é quando ocorre uma mudança abrupta no áudio, o

algoritmo pode, erroneamente, detectar a voz do narrador como o som ambiente.

Algumas abordagens audiovisuais usam tecnologia de compressão de dados e foram empregadas em vídeos com âmbito médico, esporte e independente de gênero. No primeiro caso, Cao et al. (2004) realiza segmentação de cenas em vídeos colonoscópicos, os quais são essenciais para detectar estágios iniciais de câncer no intestino. A avaliação incluiu também o domínio de vídeos sem compressão empregando técnicas distintas. Os resultados foram semelhantes em termos de detecção de cenas, mas o domínio comprimido (MPEG-2) leva um terço do tempo para processar o vídeo. No segundo caso, o algoritmo de inteligência artificial SVM auxilia a detecção de eventos em vídeos (MPEG-2) de jogos de futebol, utilizando como dados relações temporais, movimentos de câmera e descrições de tomadas. Por fim o último caso, a exemplo do anterior, utiliza não somente vídeo comprimido (MPEG), mas também faz uso do mesmo algoritmo de aprendizado de máquina, o SVM. O emprego desse algoritmo obteve melhores resultados para detectar gêneros entrevistas e piores para noticiários.

## 4.5 Características Audiovisuais com texto

O uso do texto, juntamente com as mídias de áudio e vídeo, está auxiliando pesquisadores a determinar um nível semântico melhor devido a sua capacidade de obter informação mais precisa contida no tema da cena (MANZATO; GOULARTE, 2008). Como a quantidade de mídias envolvidas no processo de segmentação é maior, métodos de unir tais mídias começam a ficar relevantes. Snoek, Worring e Smeulders (2005) apresentam duas técnicas que usam aprendizado de máquina supervisionado para unir os dados contidos nas mídias existentes: *Early Fusion* e *Late Fusion*.

A primeira integra as características de cada mídia antes dos conceitos da aprendizagem e a segunda reduz os conceitos da aprendizagem das características de cada mídia separadamente, para depois integrá-los em novos conceitos de aprendizagem, ou seja, a abordagem *Late Fusion* trata as características das mídias de maneira separada e possui mais de uma etapa no aprendizado de máquina. Essa última abordagem possui a desvantagem de ser mais complexa e cara computacionalmente dada a quantidade consideráveis de parâmetros. Como algoritmo de aprendizado, Snoek, Worring e Smeulders (2005) usa a Máquina de Vetor de Suporte (SVM), além de aplicar a abordagem em vídeos do gênero noticiário.

Wang et al. (2008) foram outros autores que fizeram uso dessa mesma abordagem mas aplicados em gêneros de programas de televisão (entretenimento). A exemplo do trabalho anterior, SVM também foi o algoritmo de inteligência artificial empregado. Dificuldades relacionadas a complexidade da técnica *Late Fusion* também foram mencionadas. Ambos os trabalhos avaliaram as duas técnicas e os melhores resultados foram encontrados na aplicação do *Late Fusion*.

O uso de SVMs também são relatadas na pesquisa de Pao et al. (2008). Contudo, esse algoritmo faz parte apenas de parte do processamento do áudio na metodologia abordada. Dentre as características relevantes desse estudo está a utilização de palavras, na mídia do texto, para navegar no conteúdo das cenas. Metodologia semelhante também foi descrita por Liu, He e Zhang (2007), os quais adotam palavras extraídas do texto e também técnicas de reconhecimento de voz para indexar o conteúdo, possibilitando a recuperação das cenas por intermédio de palavras chaves. O gênero noticiário foi empregado em ambos os trabalhos.

## 4.6 Outras Abordagens

Ao contrário dos métodos convencionais de sistemas baseados em anotações, o padrão MPEG-7 especifica meios de se fornecer informações semânticas descrevendo características audiovisuais do conteúdo multimídia. Lee, Lee e Kim (2003) descreve uma maneira de gerar índices não apenas com segmentos preliminares do filme mas também acesso não linear por meio de figuras em miniaturas . A Figura 4.2 apresenta a estrutura hierárquica utilizando a semântica do MPEG-7.

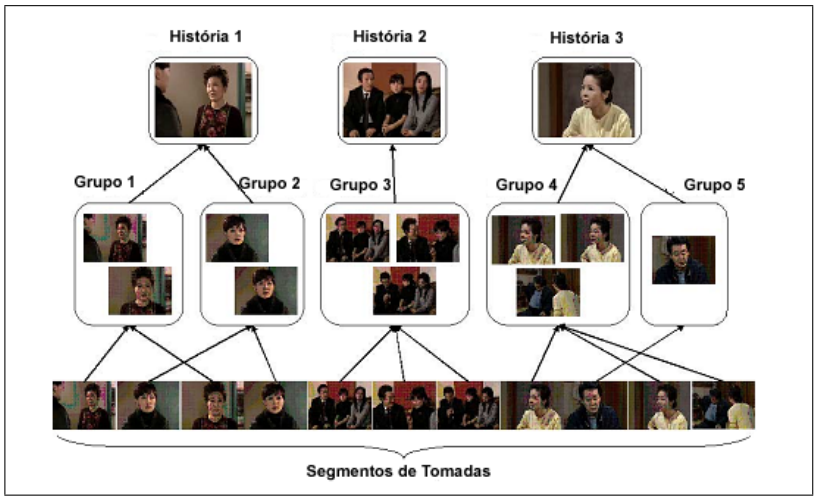



Figura 4.2: Estrutura hierárquica de representação de cena com MPEG-7 (LEE; LEE; KIM, 2003)

Contudo, não é padronizado um modo de extrair informações de alto nível do conteúdo audiovisual, necessitando de mais esforços no sentido de melhorar o nível semântico obtidos pelas técnicas de extração (VETRO; TIMMERER, 2005; BERTINI et al., 2006).

O uso de ontologias é outra abordagem recente para representar dados multimídia de uma maneira mais organizada, visando recuperação semântica facilitada. De acordo com Neto (2006) ontologia é um vocabulário consensual de termos que modela de maneira formal e abstrata um domínio de conhecimento. No contexto dessa pesquisa, o domínio de conhecimento é definido como características de imagens. Os trabalhos que envolvem

extração de informação de alto nível no domínio de imagens em ontologias são separados em dois grupos: os que definem o modelo de dados de acordo com o conteúdo multimídia (características de baixo nível) e os que modelam os dados de acordo com rótulos ou categorias semânticas atribuídas para cada imagem, como por exemplo praia, cidade, natureza, etc.

No primeiro grupo, Liu et al. (2007) relata que descritores de cor ou textura fornecem modelo de dados que facilitam a recuperação de informação semântica, como por exemplo atribuir os dados: uniforme e região azul como sendo um objeto céu. No segundo, uma ontologia utilizando categorias semânticos pré-definidas (Figura 4.3) auxilia o usuário no sentido de permitir que esse possa selecionar facilmente palavras-chaves para formular uma busca (FAN et al., 2008; FAN; GAO; LUO, 2008).

Ainda, alguns autores estão fazendo esforços para aproximar o padrão de descrição de informação multimídia, MPEG-7, ficar mais próximo linguagens de ontologias como RDF (do inglês, *Resource Description Framework*) e OWL (do inglês, *Ontology Web Language*) (HARE et al., 2006).

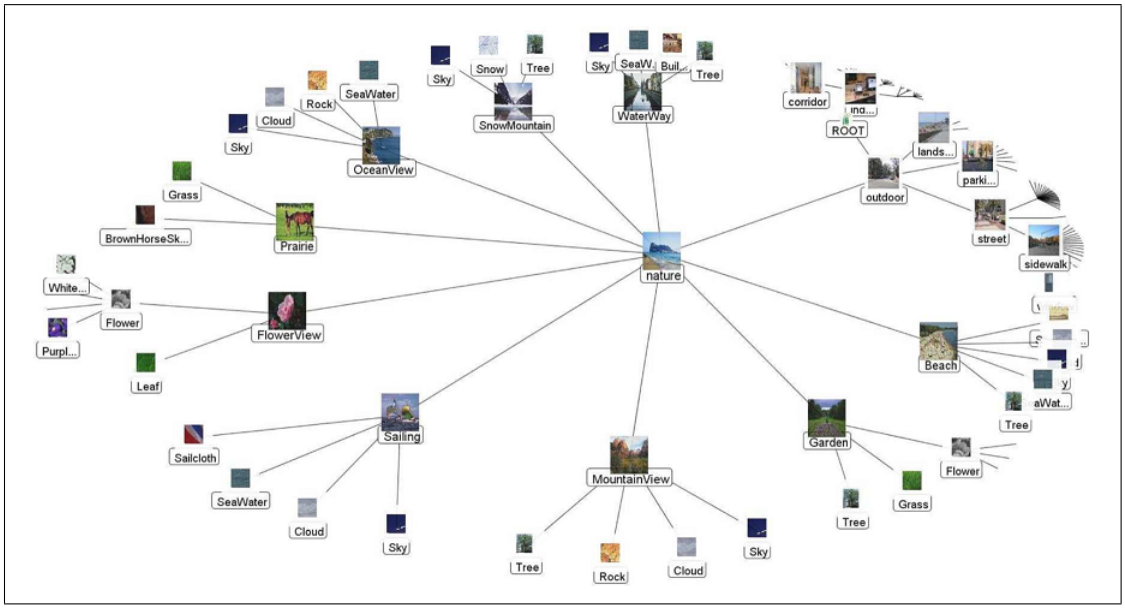



Figura 4.3: Visualização da ontologia com categorias pré-definidas (FAN et al., 2008)

Tecnologias de compressão mais sofisticadas, como o MPEG-4 (STANDARDISATION, 2002), também possui estudos na área de segmentação semântica. (CAVALLARO; STEIGER; EBRAHIMI, 2003) desenvolveu um algoritmo de transcodificação automática de conteúdo de vídeo que suporte múltiplos objetos e suas descrições. A semântica envolvida nesse estudo está relacionada a detecção de movimentação, especificamente a separação de objetos em movimento do plano de fundo, e aos descritores extraídos dos objetos de vídeo.

Outra abordagem para a extração de semântica considera a interação com o usuário como o caso da Resposta por Relevância (do inglês, *Relevance Feedback*). Essa extração

acontece por intermédio de algoritmos que tentam processar as intenções do usuário em tempo real. A medida que o usuário escolhe imagens de acordo com uma determinada busca, algoritmos de aprendizado de máquina captam essa escolha e tentam aprender com a resposta do usuário. Realocação dinâmica de pesos nas características de baixo nível pode ser efetuada quando o usuário realiza a interação (LIU et al., 2007; DEB; ZHANG, 2004).

A extração de texto de vídeos é uma metodologia que pode ajudar na classificação de vídeos. (MANZATO; GOULARTE, 2008) realiza a comparação de técnicas como algoritmos genéticos e índice semântico de latência (do inglês, *latent semantic indexing*) para determinar qual fornece o melhor resultado na classificação de vídeos de noticiários que possuam texto no formato de *closed-captions*. Apesar de LSI ser amplamente aplicada em recuperação de informação, seus resultados foram piores que os obtidos com algorítimos genéticos. Dentre as razões está o suporte a polissemia [2], que origina falsos positivos na classificação e o pequeno volume de texto empregado na amostra da metodologia.

Ferramentas que fazem uso de algoritmos de aprendizado máquina em ambas as categorias, supervisionado e não supervisionado, também conseguem obter um nível semântico mais avançado (LIU; HE; ZHANG, 2007). Na categoria supervisionado, SVM, classificador Bayesiano (JIN; SHI; CHUA, 2004) redes neurais (TOWN; SINCLAIR, 2001) e árvores de decisão (SETHI; COMAN, 2001) são utilizados para prever categoria semântica a partir de um conjunto de entrada. Erros de classificação durante a fase de treinamento e o fato de serem computacionalmente caros são as suas principais restrições. Contudo, os algoritmos não supervisionados, como *k-means* (BILENKO; BASU; MOONEY, 2004) e Corte Normalizado (do inglês, *Normalized Cut*- NCut) (NG; JORDAN; WEISS, 2002), fornecem, de modo geral, melhores resultados que os supervisionados, pois tendem a agrupar funcionalidades por semelhança, diminuindo as diferenças entre os dados de um mesmo grupo. Bons resultados em recuperação de imagens baseada em conteúdo são obtidos usando a teoria de Bayes (VASCONSELOS, 2004) para classificação por probabilidade.

## 4.7 Considerações Finais

Como apresentado neste capítulo, as pesquisas nos últimos anos estão adotando diferentes abordagens em busca de um resultado mais eficiente na área de segmentação de cenas. O processamento de características específicas do vídeo, como as visuais, áudio, texto ou combinação delas, são as principais maneiras de inferir conteúdo semântico no vídeo digital. Geralmente, os trabalhos especificam um determinado gênero do vídeo para aplicar as suas respectivas metodologias, tornando mais fácil a extração de segmentos semânticos. Tecnologias envolvendo o formato, compressão e/ou descrição dos dados do vídeo fazem parte de algumas técnicas apresentadas.

[2] Polissemia são palavras que possuem significados diferentes.

Ferramentas que, de fato, realizem recuperação confiável de conteúdo multimídia para usuários finais ainda é uma necessidade a ser atendida, tanto pela área comercial quanto pela acadêmica. Contudo, as pesquisas avançam em direção ao estreitamento da lacuna semântica ocasionada pela recuperação de informação em vídeos digitais.

Capítulo

5

# Proposta do Trabalho

## 5.1 Considerações Iniciais

Neste capítulo são descritos os detalhes e as etapas da proposta deste trabalho de mestrado. Assim, na Seção 5.2, são apresentados os objetivos. Na seção 5.3, a metodologia adotada é discutida por intermédio de uma técnica, apresentando a descrição do trabalho a ser efetuado. Na seção 5.4 o cronograma a ser seguido das atividades no desenvolvimento do trabalho é apresentado. Os resultado esperados do trabalho de pesquisa são descritos na seção 5.5. Por fim, na seção 5.6, algumas considerações finais sobre a proposta do trabalho são relatadas.

## 5.2 Objetivo da Proposta

Este plano de pesquisa tem como objetivo desenvolver um método de segmentação de vídeo em cenas baseado na semântica contida no vídeo. Um dos problemas a serem estudados é a lacuna entre a interpretação do usuário para o conteúdo da imagem e a interpretação que o computador possui ao extrair os dados dessa imagem. A segmentação de vídeo em partes semanticamente relacionadas tem por finalidade estreitar essa lacuna no gerenciamento de conteúdo multimídia. A partir disso, a adaptação de conteúdo seria beneficiada, facilitando o provimento de conteúdo personalizado levando em conta as restrições dos dispositivos portáteis, da infra-estrutura computacional do ambiente e atendendo às expectativas e preferências do usuário.

## 5.3 Metodologia

Para que os objetivos sejam alcançados, será realizada uma pesquisa aplicada, qualitativa, exploratória e experimental a respeito de técnicas de segmentação de vídeo com semântica ou identificação de cenas em vídeo. A aplicabilidade do sistema está relacionada a resolução dos desafios existentes em domínios específicos tais como aplicações que possuam conteúdo multimídia, em especial, TV Interativa e aprendizado eletrônico. Por ser qualitativa e exploratória, realizar-se-á a pesquisa por meio de estudos e comparações entre ferramentas, sistemas e aplicações. A experimentação consistirá em levantar fatores e variáveis que influenciam os resultados obtidos a partir dos métodos analisados. Durante a execução da pesquisa, o método indutivo será adotado, ou seja, as constatações obtidas a partir das experimentações, auxiliarão a elaboração de generalizações.

A metodologia do presente trabalho está fundamentada no desenvolvimento de uma técnica elaborada a partir das definições, revisões e trabalhos relacionados discutidos nos capítulos anteriores. Conforme já mencionado, a tendência das aplicações que extraem dados do vídeo é utilizar a maior quantidade de informações possível de seu conteúdo (LI; NARAYANAN; KUO, 2004; HANJALIC, 2004; NGO; ZHANG; PONG, 2001), incluindo as mídias de áudio, visual e de texto. Assim, aplicações multimodais representam o estado da arte para extração de segmentos de vídeo, especialmente os que são compostos por segmentos que possuem relação semântica entre si (cenas). A Figura 5.1 representa a técnica sugerida baseada em abordagem multimodal, detalhada a seguir.

Inicialmente, faz-se necessário a identificação das tomadas que estão presentes no fluxo de vídeo. Como o foco do estudo não é identificar tomadas, mas sim os segmentos posteriores - cenas -, essa identificação ocorrerá por meio de uma técnica já existente na literatura, visto que a detecção de tomadas é uma linha de pesquisa consolidada (ZHU; LIU, 2008a; HARB; CHEN, 2006), ou por meio de identificação manual. Após, serão extraídas características de baixo nível audiovisuais referentes a cada grupo de tomada. Pretende-se definir a quantidade e quais as características utilizadas a partir do estudo de trabalhos relacionados, no entanto, áudio, cor, textura, forma e movimentação são as mais comuns e, consequentemente, com técnicas mais consolidadas (HANJALIC, 2004). Para a extração de características visuais, quadros-chaves poderão ser utilizados para representar um grupo de tomadas.

A mídia de texto capturada deverá estar no formato de legendas, *closed-captions* ou anotações. Embora Manzato e Goularte (2008), Wang et al. (2008) explorem as informações textuais, juntamente com técnicas LSA para classificar vídeos, pretende-se adotar abordagem semelhante para classificar cenas, explorando a similaridade semântica que as informações textuais provêm entre esses segmentos para facilitar a recuperação de conteúdo com o uso de palavras-chave, por exemplo. Nessa abordagem deverão ser exploradas técnicas de aprendizado de máquina visto que muitos dos trabalhos recentes empregam

o seu uso (4.4, 4.5 e 4.6. Uma alternativa a ser explorada é utilizar o uso de anotações como informação textual, pois não há pesquisas na literatura que abordem a segmentação semântica do vídeo extraindo anotações de usuários. Por se tratar de uma ação direta do usuário no conteúdo audiovisual, essa ação pode resultar em informações semânticas com mais significado, conforme observado em MANZATO, COIMBRA e GOULARTE (2009).

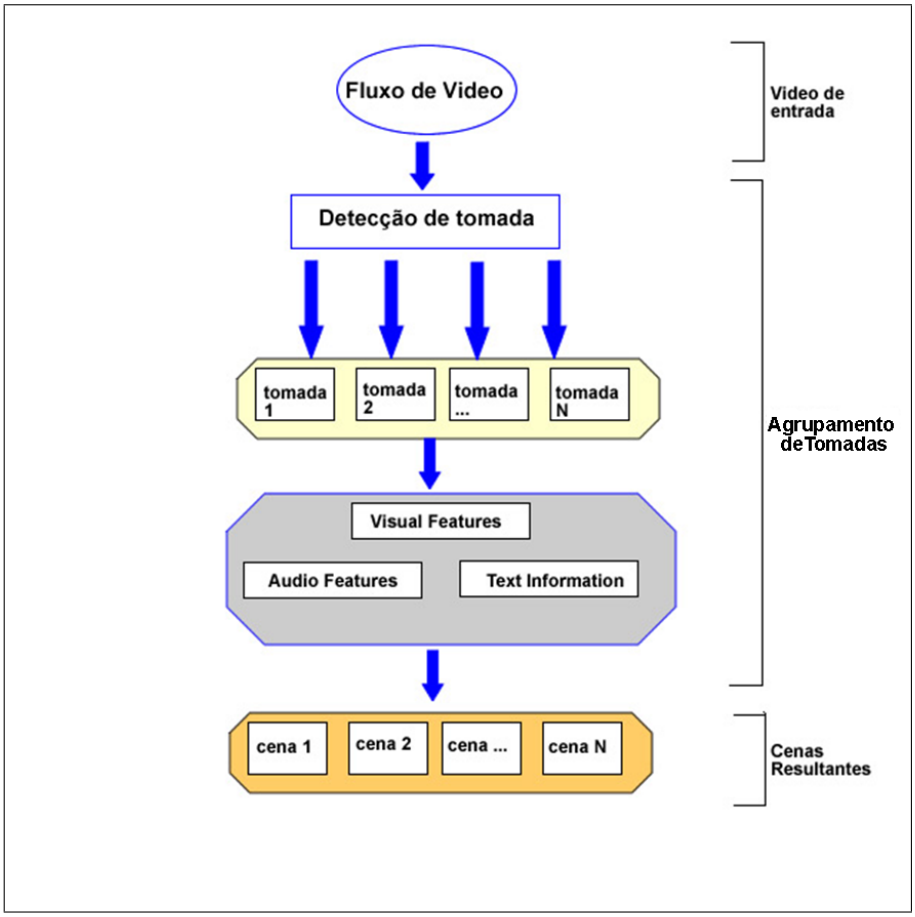



Figura 5.1: Representação da técnica proposta

Por fim, deve-se investigar um modo de agrupar e comparar os resultados das técnicas empregadas nas mídias visual e de áudio para detectar similaridades na detecção de cenas. Caso haja similaridades entre os resultados, denotando a transição entre uma cena e outra num determinado intervalo de tempo, a possibilidade de que existam duas cenas, uma antes e outra após esse intervalo, é maior devido ao fato de técnicas distintas acusarem resultado similar. Métodos que agrupam resultados similares podem ser adotados com o uso de algoritmos não supervisionados de aprendizagem de máquina (seção 4.6).

A título de exemplo, imagina-se um cenário em que uma determinada técnica detecta uma transição de cenas entre os instantes 45 e 50 segundos e outra técnica entre os instantes 42 e 47, a possibilidade de haver uma mudança de cenas entre os instantes 45 e 47 segundos é alta. Nesse contexto, as informações textuais podem ajudar a identificar essa transição caso o conjunto de palavras extraídas anteriormente a 45 segundos e poste-

riormente a 47 segundos tenham significados distintos. Ainda, os textos podem auxiliar a classificar o assunto retratado nas duas cenas e facilitar o acesso e navegação do conteúdo do vídeo. O importante nessa etapa final é usar os resultados das técnicas das mídias visual, áudio e texto de modo complementar, confirmando, ou não, a transição de cenas.

O domínio de formato com compressão digital será adotado para que suas vantagens possam ser exploradas. Os benefícios envoltos no padrão MPEG possibilitam a redução da complexidade computacional, tempo de processamento e quantidade de memória utilizada (NGO; PONG; CHIN, 1998; NGO; ZHANG; PONG, 2001; CALIC; IZQUIERDO, 2002). Desse modo, o padrão de codificação de MPEG será utilizado na fase de extração de características de baixo nível.

Os gêneros empregados para validação da técnica segue a linha dos mais utilizados nos trabalhos de análise semântica de vídeo e são escolhidos de acordo com os resultados extraídos da seção 3.4, ou seja, filmes, noticiários e esportes serão os gêneros adotados para validar a técnica. Ainda, a técnica seguirá abordagem semi-automática, pois caso seja adotado algoritmos de aprendizado de máquina, o usuário poderá participar do processo de treino (resposta por relevância), aumentando a eficiência do resultado. Os vídeos não serão processados em tempo real, constituindo, portanto, um método *off-line* de análise.

A fim de avaliar os resultados fornecidos pelo sistema descrito acima, métodos de avaliação serão empregados para calcular o desempenho e a eficiência das técnicas e algoritmos desenvolvidos. Conforme reportado no Capítulo 2, as métricas de avaliação mais utilizadas na análise de conteúdo multimídia são quantitativas, denominadas *precision* e *recall*. Contudo, essas medidas possuem restrições quanto ao conhecimento do conjunto da amostra de documentos e ao fato de capturarem diferentes aspectos do conjunto desses documentos. Portanto, em conjunto com as medidas de *precision* e *recall*, medidas de avaliação centrada no usuário também farão parte dos métodos de avaliação, pois podem fornecer um significado semântico maior. Um estudo de quais técnicas de avaliação centrada no usuário deverá ser efetuado.

## 5.4 Cronograma

A seguir, serão apresentadas atividades já realizadas e previstas para realizar o trabalho de mestrado. A tabela 5.1 apresenta o cronograma a ser seguido, de acordo com as atividades listadas.

1. **Integralização de créditos:** obtenção dos créditos obrigatórios, exigidos pelo programa de mestrado, cursando disciplinas.

2. **Exame de Proficiência em Inglês:** aprovado em 07/05/2008.

3. **Pesquisa Bibliográfica:** revisão na literatura de trabalhos que abordam os conceitos básicos da área, assim como os que apresentam colaborações para resolver as

questões em aberto do tema.

4. **Revisão Sistemática:** elaboração de uma revisão feita de maneira abrangente e com processos de pesquisas bem definidos que resultou na extração de dados relevantes nos trabalhos que abordavam técnicas, métodos ou algoritmos para realizar a segmentação de cena ou extração de informação semântica de vídeo.

5. **Extração de características e aplicação das técnicas:** extração de determinadas características de baixo nível e aplicação de suas respectivas técnicas para segmentar o vídeo em cenas.

6. **União e classificação das cenas:** união dos resultados obtidos na aplicação das técnicas e classificação semântica das cenas de acordo com a extração de informação textual.

7. **Testes e Avaliação dos Resultados:** testes e avaliações de desempenho efetuados nas técnicas utilizadas pelo sistema, usando para isso as duas maneiras de avaliação mencionadas na seção anterior.

8. **Monografia de Qualificação:** redação da monografia de qualificação.

9. **Submissão de Artigo:** escrita e submissão de artigo.

10. **Redação da Dissertação:** redação da dissertação do mestrado.

11. **Defesa da Dissertação:** defesa da dissertação de mestrado.

Tabela 5.1: Cronograma de Atividades

| | **2008** | | | | | **2009** | | | | | | **2010** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Atividades/Mê** | Mar - Abr | Mai - Jun | Jul - Ago | Set - Out | Nov - Dez | Jan - Fev | Mar - Abr | Mai - Jun | Jul - Ago | Set - Out | Nov - Dez | Jan - Fev |
| 1) Créditos | ■ | ■ | ■ | ■ | ■ | | ■ | ■ | | | | |
| 2) Exame - Inglês | | ■ | | | | | | | | | | |
| 3) Pesquisa Bibliográfica | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | | | | |
| 4) Revisão Sistemática | | | ■ | ■ | ■ | | | | | | | |
| 5) Execução - Técnicas | | | | | | | | ■ | ■ | ■ | | |
| 6) União - Classificação | | | | | | | | | | ■ | ■ | |
| 7) Testes e Avaliação | | | | | | | | | | | ■ | ■ |
| 8) Monografia - Qualificação | | | | | | ■ | | | | | | |
| 9) Submissão de Artigo | | | | | | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| 10) Redação da Dissertação | | | | | | | | | | | ■ | ■ |
| 11) Defesa da Dissertação | | | | | | | | | | | | ■ |

Entre os resultados esperados do presente projeto, destacam-se:

- Disponibilização de técnicas para realizar adaptação de conteúdo multimídia.

- Validação dessas técnicas por meio da implementação de uma aplicação que segmente vídeos usando em semântica.
- Publicação de resultados em conferências e periódicos relacionados.
- Atualização do conhecimento do grupo em estratégias de personalização e adaptação de conteúdo.
- Possibilidade de estender a técnica para diferentes gêneros de vídeo.
- Aplicar a técnica no domínio de personalização de conteúdo, contribuindo com o trabalho do aluno de doutorado Marcelo Manzato.

## 5.5 Considerações Finais

Este capítulo teve como intuito relatar os objetivos desse trabalho de pesquisa e propor uma abordagem para contribuir com o estado da arte na recuperação de informação de conteúdo multimídia. A metodologia sugerida foi elaborada de acordo com os conceitos, desafios e estudos relacionados ao tema do projeto.

Como resultados obtidos até o momento estão: *i*) a produção de uma revisão sistemática que teve como objetivos coletar trabalhos que contenham identificação de cenas ou semântica envolvida em segmentação temporal de vídeo digital; *ii*) um trabalho publicado no congresso EUROITV que retrata a classificação de vídeos (noticiários) utilizando ferramenta que capta as anotações do usuário no conteúdo multimídia (MANZATO; COIMBRA; GOULARTE, 2009) e *iii*) a revisão bibliográfica de trabalhos que fazem análise de vídeo digital, abordando conceitos e técnicas para extração de informação.

# Referências Bibliográficas


AL-HAMES, M. et al. A two-layer graphical model for combined video shot and scene boundary detection. In: *Proc. IEEE International Conference on Multimedia and Expo.* [S.l.: s.n.], 2006. p. 261–264.

ANER-WOLF, A.; KENDER, J. R. Video summaries and cross-referencing through mosaic-based representation. *Comput. Vis. Image Underst.*, Elsevier Science Inc., New York, NY, USA, v. 95, n. 2, p. 201–237, 2004.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. *Modern Information Retrieval.* Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

BARRIOS, V. M. G.; MöDRITSCHER, F.; GüTL, C. Personalization versus Adaptation? A User-centred Model Approach and its Application. In: *Proceedings of I-KNOW'05.* [S.l.: s.n.], 2005. p. 120–127.

BERTINI, M. et al. Semantic Adaptation of Sport Videos With User-Centred Performance Analysis. *IEEE Transactions on Multimedia*, v. 8, n. 3, p. 433–443, 2006.

BILENKO, M.; BASU, S.; MOONEY, R. J. Integrating constraints and metric learning in semi-supervised clustering. In: *Proceedings of the 21st. International Conference on Machine Learning (ICML).* [S.l.: s.n.], 2004. p. 81–88.

BIOLCHINI, J. et al. *Systematic review in software engineering: Relevance and utility.* [S.l.], 2005.

CALIC, J.; IZQUIERDO, E. Temporal segmentation of mpeg video streams. *EURASIP J. Appl. Signal Process.*, Hindawi Publishing Corp., New York, NY, United States, v. 2002, n. 1, p. 561–565, 2002.

CAO, J.-R. Algorithm of scene segmentation based on svm for scenery documentary. IEEE Computer Society, Washington, DC, USA, p. 95–98, 2007.

CAO, Y. et al. A visual model approach for parsing colonoscopy videos. In: . [S.l.: s.n.], 2004. p. 160–169.

CAVALLARO, A.; STEIGER, O.; EBRAHIMI, T. Semantic segmentation and description for video transcoding. *ICME '03: Proceedings of the 2003 International Conference on Multimedia and Expo (ICME '03)*, v. 3, p. 597–600, 2003.

CHEN, L.-H.; LAI, Y.-C.; LIAO, H.-Y. M. Movie scene segmentation using background information. *Pattern Recogn.*, Elsevier Science Inc., New York, NY, USA, v. 41, n. 3, p. 1056–1065, 2008.

CORREIA, P.; PEREIRA, F. Classification of video segmentation application scenarios. *Circuits and Systems for Video Technology, IEEE Transactions on*, v. 14, n. 5, p. 735–741, 2004.

DEB, S.; ZHANG, Y. An overview of content-based image retrieval techniques. In: *AINA '04: Proceedings of the 18th International Conference on Advanced Information Networking and Applications*. Washington, DC, USA: IEEE Computer Society, 2004. p. 59.

DIMITROVA, N. et al. Applications of video-content analysis and retrieval. *Multimedia, IEEE*, v. 9, n. 3, p. 42–55, 2002.

DONG, A.; LI, H. Semantic segmentation of documentary video using music breaks. *Multimedia and Expo, 2006 IEEE International Conference on*, p. 1825–1828, 2006.

FAN, J.; GAO, Y.; LUO, H. Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *Image Processing, IEEE Transactions on*, v. 17, n. 3, p. 407–426, 2008.

FAN, J. et al. Mining multilevel image semantics via hierarchical classification. *IEEE Transactions on Multimedia*, v. 10, n. 2, p. 167–187, 2008.

GOULARTE, R. *Personalização e adaptação de conteúdo baseadas em contexto para TV Interativa*. Tese (Doutorado) — ICMC-USP, São Carlos, 2003.

GU, Z. et al. Ems: Energy minimization based video scene segmentation. *Multimedia and Expo, 2007 IEEE International Conference on*, p. 520–523, July 2007.

HANJALIC, A. *Content-Based Analysis of Digital Video*. [S.l.]: Kluwer Academic Publishers, 2004. 193 pags.

HANJALIC, A. et al. Indexing and retrieval of tv broadcast news using dancers. *Journal of Electronic Imaging*, SPIE, v. 10, n. 4, p. 871–882, 2001.

HARB, H.; CHEN, L. Audio-based description and structuring of videos. *International Journal on Digital Libraries*, v. 6, n. 1, p. 70–81, 2006.

HARE, J. S. et al. Mind the gap: another look at the problem of the semantic gap in image retrieval. In: CHANG, E. Y.; HANJALIC, A.; SEBE, N. (Ed.). [S.l.]: SPIE, 2006. v. 6073, n. 1, p. 607309.

HOUAISS, A. *Dicionário Houaiss de Língua Portuguesa*. 1a edição. ed. Rio de Janeiro: Villar Ms, 2001.

JIANG, H.; LIN, T.; ZHANG, H.-J. Video segmentation with the assistance of audio content analysis. *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, v. 3, p. 1507–1510 vol.3, 2000.

JIN, W.; SHI, R.; CHUA, T. S. A semi-naive bayesian method incorporating clustering with pair-wise constraints for auto image annotation. In: *Proceedings of the ACM Multimedia*. [S.l.: s.n.], 2004.

JOYCE, R.; LIU, B. Temporal segmentation of video using frame and histogram space. *Multimedia, IEEE Transactions on*, v. 8, n. 1, p. 130–140, 2006.

KITCHENHAM, B. *Procedures for performing systematic reviews*. [S.l.], 2004.

KOPRINSKA, I.; CARRATO, S. Temporal video segmentation: A survey. *Signal Processing: Image Communication*, v. 16, p. 477–500(24), 2001.

LEE, J.-H.; LEE, G.-G.; KIM, W.-Y. Automatic video summarizing tool using mpeg-7 descriptors for personal video recorder. v. 49, n. 3, p. 742–749, 2003.

LI, Y.; MING, W.; KUO, C.-C. Semantic video content abstraction based on multiple cues. In: *Proc. IEEE International Conference on Multimedia and Expo ICME 2001*. [S.l.: s.n.], 2001. p. 623–626.

LI, Y.; NARAYANAN, S.; KUO, C. Content-based movie analysis and indexing based on audiovisual cues. v. 14, n. 8, p. 1073–1085, 2004.

LIU, H.; HE, T.; ZHANG, H. Nbr: A content-based news video browsing and retrieval system. In: *Technologies for E-Learning and Digital Entertainment*. [S.l.: s.n.], 2007. p. 793–800.

LIU, Y. et al. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, v. 40, p. 262–282, 2007.

LOWE, D.; HALL, W. *Hypermedia & the Web*. [S.l.]: John Wiley & Sons Ltd, 1999.

LU, L.; CAI, R.; HANJALIC, A. Audio elements based auditory scene segmentation. In: *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2006*. [S.l.: s.n.], 2006. v. 5, p. 17–20.

LUM, W. Y.; LAU, F. C. M. A Context-Aware Decision Engine for Context Adaptation. *IEEE Pervasive Computing*, v. 1, n. 3, p. 41–49, 2002.

MAGALHãES, J.; PEREIRA, F. Using mpeg standards for multimedia customization. *Signal Processing: Image Communication*, v. 19, p. 437–456, 2004.

MANZATO, M.; GOULARTE, R. Video news classification for automatic content personalization: A genetic algorithm based approach. *In Proceedings of the XIV Webmedia*, 2008.

MANZATO, M. G.; COIMBRA, D. B.; GOULARTE, R. Multimedia content personalization based on peer-level annotation. In: 7TH EUROPEAN INTERACTIVE TV CONFERENCE. *European Interactive TV Conference*. [S.l.], 2009. v. 1, p. 1–8. (to appear).

MORISAWA, K.; NITTA, N.; BABAGUCHI, N. Video scene retrieval with sign sequence matching based on audio features. In: . [S.l.: s.n.], 2005. p. 121–129.

NETO, R. F. B. *Um processo de software e um modelo ontológico para apoio ao desenvolvimento de aplicações sensíveis ao contexto*. Tese (Doutorado) — ICMC-USP, São Carlos, 2006.

NG, A. Y.; JORDAN, M. I.; WEISS, Y. On spectral clustering: analysis and an algorithm. *Advances in Neural Information Processing Systems*, MIT Press, v. 14, 2002.

NGO, C. W.; PONG, T. C.; CHIN, R. T. A survey of video parsing and image indexing techniques in compressed domain. In: *Symposium on Image and Speech and Signal Processing and Robotics*. [S.l.: s.n.], 1998. p. 231–236.

NGO, T.-w.; ZHANG, H.-j.; PONG, T.-c. Recent advances in content based video analysis. *International Journal of Image and Graphics*, v. 1, p. 1–3, 2001.

OH, J. H. et al. Video abstraction. In: DEB, S. (Ed.). *Video Data Management and Information Retrieval*. [S.l.]: Idea Group Publishing, 2005.

PAI, M. et al. Clinical research methods - systematic reviews and meta-analyses: An illustrated, step-by-step guide. *The National Medical Journal of India*, v. 17, p. 86–94, 2004.

PAO, H. T. et al. Constructing and application of multimedia tv-news archives. In: . Tarrytown, NY, USA: Pergamon Press, Inc., 2008. v. 35, n. 3, p. 1444–1450.

PORTER, S.; MIRMEHDI, M.; THOMAS, B. Temporal video segmentation and classification of edit effects. *Image and Vision Computing*, v. 21, p. 1097–1106, 2003.

RICHARDSON, I. E. G. *H.264 and mpeg-4 video compression*. [S.l.: s.n.], 2003. 315 pags.

RUI, Y.; HUANG, T.; MEHROTRA, S. Exploring video structure beyond the shots. In: *Proc. IEEE International Conference on Multimedia Computing and Systems*. [S.l.: s.n.], 1998. p. 237–240.

SETHI, I. K.; COMAN, I. L. Mining association rules between low-level image features and high-level concepts. In: *Proceedings of the SPIE Data Mining and Knowledge Discovery*. [S.l.: s.n.], 2001. v. 3, p. 279–290.

SHAO, X. et al. Automatic summarization of music videos. *ACM Trans. Multimedia Comput. Commun. Appl.*, ACM, v. 2, n. 2, p. 127–148, 2006.

SMEATON, A. F. Techniques used and open challenges to the analysis, indexing and retrieval of digital video. *Information Systems*, Elsevier Science Ltd., Oxford, UK, UK, v. 32, n. 4, p. 545–559, 2007.

SMEULDERS, A. et al. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, v. 22, n. 12, p. 1349–1380, 2000.

SNOEK, C. G. M.; WORRING, M.; SMEULDERS, A. W. M. Early versus late fusion in semantic video analysis. In: *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*. New York, NY, USA: ACM, 2005. p. 399–402.

SRINIVASAN, U.; NEPAL, S. *Managing Multimedia Semantics*. [S.l.]: IRM Press, 2005. 409 pag.

STANDARDISATION, I. O. for. *Short MPEG-2 Description*. 2000. Disponível em: <http://www.chiariglione.org/mpeg/standards/mpeg-2/mpeg-2.htm>. Acesso em: fevereiro de 2009.

STANDARDISATION, I. O. for. *MPEG-4 Description*. 2002. Disponível em: <http://www.chiariglione.org/mpeg/standards/mpeg-4/mpeg-4.htm>. Acesso em: fevereiro de 2009.

SURAL, S.; MOHAN, M.; MAJUMDAR, A. K. A soft decision histogram from the hsv color space for video shot detection. In: DEB, S. (Ed.). *Video Data Management and Information Retrieval*. [S.l.]: Idea Group Publishing, 2005.

TOWN, C. P.; SINCLAIR, D. Content-based image retrieval using semantic visual categories. *Society for Manufacturing Engineers. Technical Report*, 2001.

VASCONSELOS, N. On the efficient evaluation of probabilistic similarity functions for image retrieval. *IEEE Transactions on Inf. Theory*, v. 50, n. 7, p. 1482–1496, 2004.

VETRO, A.; TIMMERER, C. Digital Item Adaptation: Overview of Standardization and Research Activities. *IEEE Transactions on Multimedia*, v. 7, n. 3, p. 418–426, 2005.

WANG, J.; CHUA, T.-S. A cinematic-based framework for scene boundary detection in video. *The Visual Computer*, v. 19, n. 5, p. 329–341, 2003.

WANG, J. et al. A multimodal scheme for program segmentation and representation in broadcast video streams. *Multimedia, IEEE Transactions on*, v. 10, n. 3, p. 393–408, 2008.

WANG, Y.; LIU, Z.; HUANG, J.-C. Multimedia content analysis-using both audio and visual clues. v. 17, n. 6, p. 12–36, 2000.

WU, C.-H.; IRWIN, J.; DAI, F. Enabling multimedia applications for factory automation. *Industrial Electronics, IEEE Transactions on*, v. 48, n. 5, p. 913–919, 2001.

YOO, H.-W. Retrieval of movie scenes by semantic matrix and automatic feature weight update. *Expert Syst. Appl.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 34, n. 4, p. 2382–2395, 2008.

ZHAI, Y.; SHAH, M. Video scene segmentation using markov chain monte carlo. *Multimedia, IEEE Transactions on*, v. 8, n. 4, p. 686–697, 2006.

ZHANG, Y.-J. (Ed.). *Advances in Image and Video Segmentation*. Hershey, Pa, USA: IRM Press, 2006.

ZHANG, Y.-J.; JIANG, F. Home video structuring with a two-layer shot clustering approach. In: *Proc. 3rd International Symposium on Communications, Control and Signal Processing ISCCSP 2008*. [S.l.: s.n.], 2008. p. 500–504.

ZHAO, L.; YANG, S.-Q.; FENG, B. Video scene detection using slide windows method based on temporal constrain shot similarity. In: *Proc. IEEE International Conference on Multimedia and Expo ICME 2001*. [S.l.: s.n.], 2001. p. 1171–1174.

ZHU, S.; LIU, Y. Automatic scene detection for advanced story retrieval. *Expert Systems with Applications*, In Press, Uncorrected Proof, 2008.

ZHU, S.; LIU, Y. A novel scheme for video scenes segmentation and semantic representation. In: *Proc. IEEE International Conference on Multimedia and Expo*. [S.l.: s.n.], 2008. p. 1289–1292.

ZUTSCHI, S. et al. Managing multimedia semantics. In: _____. [S.l.]: Idea Group Inc., 2005. cap. The role of elevance feedback in managing multimedia semantics: A survey, p. 288–304.

# Glossário

**Árvore de Decisão** – É um dos métodos de aprendizado simbólico mais amplamente utilizados e práticos para inferência indutiva. É utilizado para aproximar funções discretas robustas a dados com ruído e que permite o aprendizado de expressões disjuntas. Este método de aprendizagem está entre os mais populares algoritmos de inferência indutiva e foi aplicado amplamente em diferentes domínios.

**Codificação** – Em processamento digital de sinais, codificação significa a modificação de características de um sinal para torná-lo mais apropriado para uma aplicação específica, como por exemplo, compressão, transmissão ou armazenamento de dados.

**Classificador Bayesiano** – É um sistema de aprendizado supervisionado que, utilizando-se da teoria de Redes Bayesianas, é capaz de combinar informações para elaborar hipóteses. A Rede Bayesiana é baseada na teoria da probabilidade e é empregada para o tratamento do conhecimento a partir da incerteza através de inferência. As regras de decisão encontradas pela rede bayesiana determinam a hipótese mais provável dado o conhecimento e as evidências disponíveis na rede.

**Decodificação** – Decodificação é o processo contrário da codificação, ou seja, as características modificadas de um sinal são transformadas em seu formato original.

**Dispositivos Portáteis** – São dispositivos que podem ser utilizados em qualquer lugar e a qualquer hora. São caracterizados pelo pequeno porte, e geralmente são projetados para atender uma certa funcionalidade. Exemplos de dispositivos portáteis são o telefone celular, PDA, *tablet*, etc.

**Estimativa e Compensação de Movimento** – Técnica utilizada para eliminar a redundância temporal de um sinal de vídeo. O algoritmo compara quadros próximos uns dos outros e verifica quais áreas da imagem são equivalentes ou mesmo qual foi o movimento realizado pela área de um quadro para outro. Essas áreas equivalentes são eliminadas, e

no lugar, são armazenadas no arquivo codificado apenas essas informações de controle, que serão mais tarde processadas pelo decodificador, para reconstrução da imagem original.

**Fluxo** – Em computação, refere-se a um fluxo de dados.

**K-Means** – É considerado um algoritmo de mineração de dados não-supervisionado que fornece uma classificação de informações de acordo com os próprios dados. Essa classificação é baseada em análise e comparações entre os valores numéricos dos dados. Assim, o algoritmo automaticamente fornece uma classificação sem a necessidade de supervisão humana, ou seja, sem nenhuma pré-classificação existente.

**Largura de Banda** – Intervalo do espectro de frequências disponível ou necessário para transmitir dados (imagens, áudio, pacotes digitais) sobre um meio, tal como cabo ou ar, ou sobre um dispositivo elétrico. Quanto maior é a largura de banda disponível, maior é a quantidade de dados que pode ser transmitida por segundo.

**Multimídia** – É a utilização simultânea de vários tipos de mídia (texto, sons, imagens, gráficos, vídeos e animações).

**Ontologia** – Em computação, o termo ontologia tem como princípio básico: o que "existe" é o que pode ser representado. Nesse contexto, ontologia pode ser entendida como uma especificação formal e explícita de uma conceitualização consensual, a qual pode ser definida como uma estrutura composta por um domínio de conhecimento e um conjunto de relações sobre o mesmo.

**Personalização** – É o processo no qual um sistema se adapta a fim de satisfazer os requisitos de determinado usuário.

**Quadro de Vídeo** – É uma das inúmeras imagens que compõem um vídeo; ao tocar um vídeo, cada quadro é mostrado na tela por um tempo especificado pela taxa de quadros: se o vídeo está configurado a 25 quadros/s, então cada quadro será apresentado por um período de $0,040$ segundos.

**Redes Neurais** – São sistemas computacionais baseados numa aproximação à computação baseada em ligações. Nós simples são interligados para formar uma rede de nós – daí o termo "rede neural". A inspiração original para essa técnica provém do exame das estruturas do cérebro, em particular do exame de neurônios.

**SVM** – *Support Vector Machine* (Máquina de Vetor de Suporte). É definido como um conjunto de métodos de aprendizagem supervisionados usados para classificação e regressão. Uma propriedade especial de SVMs é que eles simultaneamente minimizam o erro de classificação empírica e maximizam a margem geométrica.

**TV Digital** – É quando sinais de televisão são devolvidos em uma forma digital. As

vantagens da TV Digital são o aumento da qualidade e a largura de banda reduzida. Além disso, difusão digital permite o desenvolvimento de serviços de TV Interativa.

**UMA** – *Universal Multimedia Access* (Acesso Multimídia Universal). Conceito referente ao acesso a informações multimídia independentemente do dispositivo ou rede utilizados. O objetivo é disponibilizar diferentes formatos de um mesmo conteúdo personalizados para cada situação de rede, dispositivo ou preferências de usuário.

**Vetor de Movimento** – Termo utilizado em compressão de vídeo. Indica a translação espacial de um bloco para outro em quadros distintos, onde essa translação é especificada pela aplicação da técnica estimativa de movimento.