
**Segmentação de cenas em telejornais:
uma abordagem multimodal**

Danilo Barbosa Coimbra

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:

Segmentação de cenas em telejornais: uma abordagem multimodal

Danilo Barbosa Coimbra

Orientador: *Prof. Dr. Rudinei Goularte*

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. *VERSAO REVISADA.*

USP – São Carlos
Junho/2011

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

C679s Coimbra, Danilo Barbosa
Segmentação de cenas em telejornais: uma abordagem
multimodal / Danilo Barbosa Coimbra; orientador
Rudinei Goularte -- São Carlos, 2011.
107 p.

Dissertação (Mestrado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2011.

1. Segmentação de cena. 2. Segmentação de vídeo. 3.
Identificação de cena. 4. Identificação de transição de
cenas. I. Goularte, Rudinei, orient. II. Título.

Dedicatória

*Aos meus pais Moacir e Maria Helena, por terem
me dado a luz da vida e amor incondicional.*

Aos meus irmãos Karina e Hugo, pela valiosa amizade.

*Aos meus amigos, por terem colaborado
com minha formação pessoal e profissional.*

Agradecimentos

Inicialmente, agradeço aos meus queridos pais por todo o esforço que fizeram para que eu pudesse atingir os meus objetivos, sejam eles pessoais, acadêmicos e profissionais. Obrigado por me amar incondicionalmente, acreditar e confiar em mim sempre!

Agradeço aos meus inestimáveis irmãos, Karina e Hugo, que em meio a desentendimentos, estes poucos significaram diante do carinho, afeto e amizade que compartilhamos. Os momentos que tivemos e que teremos são provas da nossa união.

Agradeço aos meus amigos e companheiros de São Carlos, em especial, aos do ICMC-USP pela ajuda, incentivo e companhia tanto nos momentos de estudo quanto nos momentos de lazer: Marcelo Manzato, Willian “watinha” Watanabe, Vivian Motti, David “Conan” Fernandes, José “Tim” Costa Martins, Adalberto “Daga” Gonzaga, Roberto Rigolin, Roberto Sadao, Diogo Predosa, Eduardo “Belezuti” Pezutti, Gabriel “Ceará” Queiroz, Daniel Lobato, Rafael “Frotinha”, Cassio Prazeres.

Agradeço aos meus amigos-irmãos Jonas Migano, Humberto Costa e Vera, Gabriel Tarquinio, por todos os momentos em que estamos juntos. Aprendo muito com vocês.

Agradeço aos meus amigos de “república” Maycon Leone e Tácito Trindade, pelos momentos de descontração, apoio e incentivo, muitas vezes estudando madrugada a dentro tomindo litros de café com o Tácito ou ouvindo as risadas do Maycon.

Agradeço à minha companheira Flavia, por me proporcionar sentimentos únicos de carinho, afeto e amor, não esquecendo da paciência e compreensão que foram essenciais para o término desta jornada.

Agradeço aos ex-companheiros de trabalho da antiga empresa GFMI (agora com o nome de Ailog), os quais mesmo distantes, sempre torceram por mim.

Agradeço à todos os funcionários e professores do instituto pelo suporte e profissionalismo sempre presente.

Agradeço àqueles que fizeram e fazem parte da minha vida, pois colaboraram de modo direto ou indireto para a finalização desta jornada.

Por último, porém não menos importante, quero agradecer ao meu orientador Prof. Rudinei Goularte, pelas suas conversas, conselhos, orientações e sugestões, sem os quais não seria possível o início e o término deste trabalho. Obrigado!

Sumário

Lista de Figuras	vii
Lista de Tabelas	ix
Lista de Siglas	xi
1 Introdução	1
2 Conceitos, Definições e Metodologias	5
2.1 Considerações Iniciais	5
2.2 Estrutura do vídeo digital	6
2.3 Análise do Vídeo Digital	9
2.3.1 Detecção de Tomadas	10
2.3.2 Detecção de Quadros-Chave	13
2.3.3 Detecção de Cenas e a Lacuna Semântica	14
2.3.4 Cenas em Telejornais	17
2.4 Métodos de Segmentação de Vídeo	19
2.4.1 Histogramas de Cor	20
2.4.2 Transformada Discreta de Wavelet	21
2.4.3 <i>Root Mean Square</i>	24
2.5 Avaliação de Resultados	25
2.6 Considerações Finais	26
3 Segmentação de Cenas em Vídeo Digital	29
3.1 Considerações Iniciais	29
3.2 Características Visuais	30
3.3 Características de Áudio	33
3.4 Características Multimodais	35
3.4.1 Características Audiovisuais	35

3.4.2	Características Visuais e Textuais	36
3.4.3	Características Audiovisuais com Texto	38
3.5	Outras Abordagens	41
3.6	Considerações Finais	43
4	Segmentação Multimodal de Cenas: uma proposta	45
4.1	Considerações Iniciais	45
4.2	Técnicas Empregadas	45
4.2.1	Extração de Informação em Imagens	46
4.2.2	Extração de Informação em Áudio	47
4.2.3	Extração de Informação em Texto	49
4.3	Técnica Multimodal Proposta	51
4.4	Resultados Obtidos	55
4.4.1	Ambiente de Testes	57
4.4.2	Técnica Textual	58
4.4.3	Técnicas Sonoras	59
4.4.4	Técnicas Visuais	61
4.4.5	Técnica Multimodal	65
4.5	Considerações Finais	74
5	Conclusões e Trabalhos Futuros	79
5.1	Considerações Iniciais	79
5.2	Contribuições	79
5.3	Limitações	80
5.4	Trabalhos Futuros	80
Referências Bibliográficas		83
Apêndice		95
Glossário		103

Lista de Figuras

2.1	Algoritmos aplicados nos vídeos digitais que são desenvolvidos para facilitar a recuperação de seus conteúdos. Adaptado de Hanjalic (2004)	7
2.2	Estrutura do fluxo de vídeo digital	7
2.3	Outras representações para estrutura de vídeo digital	8
2.4	Transição abrupta de tomada	11
2.5	Dissolução (Porter et al., 2003)	12
2.6	Transição <i>fade out</i> seguida por <i>fade in</i> (Koprinska & Carrato, 2001)	12
2.7	Diferentes tipos de <i>wipes</i> (Joyce & Liu, 2006)	12
2.8	Representação da lacuna semântica. Adaptado de (Hanjalic, 2004)	15
2.9	Pirâmide da estrutura do conteúdo do vídeo. Traduzido de (Hanjalic, 2004)	16
2.10	Modelo de indexação semântica hierárquica de vídeo proposto por Snoek & Worring (2005)	17
2.11	Composição temporal dos telejornais	18
3.1	Âncoras com planos de fundo dinâmicos	33
3.2	Visão geral dos métodos de detecção de cena com áudio. Traduzido de Jiang et al. (2000)	34
3.3	Arquitetura da técnica de união das características visuais e de áudio (Coimbra & Goularte, 2009)	36
3.4	Estrutura hierárquica de representação de cena com MPEG-7 (Lee et al., 2003)	41
3.5	Visualização da ontologia com categorias pré-definidas (Fan et al., 2008b) .	42
4.1	Imagem modelo do âncora para busca	46
4.2	Parte da Interface visual da ferramenta Audacity	48
4.3	Representações do <i>closed-caption</i>	49
4.4	Arquitetura da técnica multimodal	52
4.5	Exemplo de um <i>closed-caption</i> mal formatado	59
4.6	Imagem modelo para busca no Jornal Nacional de 09/03/2010	63

4.7	Imagens resultantes da técnica de histograma global para o Jornal Nacional de 09/03/2010	63
4.8	Imagens resultantes da técnica de wavelet para o Jornal Nacional de 09/03/2010	64
4.9	Imagens resultantes da técnica de histograma local para o Jornal da Record de 23/07/2010	65
4.10	Imagens resultantes da técnica de wavelet para o Jornal Nacional de 09/03/2010	65
4.11	Jornal Nacional 22/02/2010	67
4.12	Jornal Nacional 02/03/2010	67
4.13	Jornal Nacional 03/03/2010	68
4.14	Jornal Nacional 04/03/2010	68
4.15	Jornal Nacional 09/03/2010	69
4.16	Jornal Hoje 02/03/2010	69
4.17	Jornal Hoje 03/03/2010	70
4.18	Jornal Hoje 05/03/2010	70
4.19	Jornal Hoje 06/03/2010	71
4.20	Jornal Hoje 10/03/2010	71
4.21	Jornal da Record 16/07/2010	72
4.22	Jornal da Record 19/07/2010	72
4.23	Jornal da Record 20/07/2010	73
4.24	Jornal da Record 21/07/2010	73
4.25	Jornal da Record 23/07/2010	74
4.26	Jornal da Globo 04/03/2010	74
4.27	Jornal da Globo 08/09/2010	75
4.28	Jornal da Globo 09/09/2010	75
4.29	Jornal da Globo 21/09/2010	76
4.30	Jornal da Globo 23/09/2010	76
5.1	Distribuição do número de artigos por periódico	97
5.2	Distribuição da quantidade de artigos pelo tempo	98
5.3	Porcentagem de produção no decorrer do tempo	98
5.4	Distribuição do número de artigos após fase de seleção por resumo	99
5.5	Distribuição da quantidade de artigos pelo tempo, após fase de seleção por resumo	99
5.6	Gêneros de vídeos identificados no processo de validação das técnicas	100
5.7	Quais tipos de mídias foram extraídas em cada trabalho	100
5.8	Quantidade dos trabalhos que utilizam um formato de compressão(em %) .	101

Lista de Tabelas

2.1	Tabela de algumas características de baixo-nível com suas respectivas técnicas (Hanjalic, 2004)	14
3.1	Tabela das principais características textuais com suas respectivas vantagens e desvantagens (adaptado de Brezeale & Cook (2008))	37
4.1	Resultados das técnicas	56
4.2	Informações da estrutura temporal dos telejornais	57
4.3	Resultados da técnica de texto	58
4.4	Resultados das técnicas sonoras	60
4.5	Resultados da identificação das transições de propagandas publicitárias com as técnicas de áudio	61
4.6	Resultados das técnicas visuais	62
4.7	Resultados da técnica multimodal	66

Lista de Siglas

- ACM – *Association for Computing Machinery*
ASR – *Automatic Speech Recognition*
CBVR – *Content-Based Video Retrieval*
CBIR – *Content-Based Image Retrieval*
CC – *Closed-Caption*
CWT – *Continuous Wavelet Transform*
dB – *Decibéis*
DCT – *Discrete Cossin Transform*
DWT – *Discrete Wavelet Transform*
FWT – *Fast Wavelet Transform*
GHT – *Generalized Hough Transform*
HSV – *Hue Saturation Value*
HMM – *Hidden Markov Models*
ICMC – Instituto de Ciências Matemáticas e de Computação
IEEE – *Institute of Electrical and Electronics Engineers*
kHz – *Kilo Hertz*
LSI – *Latent Semantic Indexing*
LSA – *Latent Semantic Analysis*
MCMC – *Markov Chain Monte Carlo*
MDCT – *Modified Discrete Cosine Transform*
MPEG – *Moving Picture Experts Group*
MP3 – *MPEG 1 Layer-3*
NCut – *Normalized Cut*
OCR – *Optical Caption Recognition*
OWL – *Ontology Web Language*
P&A – Personalização e Adaptação de Conteúdo
RGB – *Red-Green-Blue*

RMS – *Root Mean Square*

RDF – *Resource Description Framework*

SNR – *Signal-to-noise ratio*

SRT – SubRip (formato de legenda)

SVM – *Support Vector Machine*

TV – Televisão

USP – Universidade de São Paulo

UMA – *Universal Multimedia Access*

ZCR – *Zero-Crossing Rate*

WWW – *World Wide Web*

Resumo

Este trabalho tem como objetivo desenvolver um método de segmentação de cenas em vídeos digitais que trate segmentos semânticamente complexos. Como prova de conceito, é apresentada uma abordagem multimodal que utiliza uma definição mais geral para cenas em telejornais, abrangendo tanto cenas onde âncoras aparecem quanto cenas onde nenhum âncora aparece. Desse modo, os resultados obtidos da técnica multimodal foram significativamente melhores quando comparados com os resultados obtidos das técnicas monomodais aplicadas em separado. Os testes foram executados em quatro grupos de telejornais brasileiros obtidos de duas emissoras de TV diferentes, cada qual contendo cinco edições, totalizando vinte telejornais.

Palavras-chave: segmentação de cenas, segmentação semântica, identificação de cenas, identificação de transição de cenas, técnica multimodal.

Abstract

This work aims to develop a method for scene segmentation in digital video which deals with semantically complex segments. As proof of concept, we present a multimodal approach that uses a more general definition for TV news scenes, covering both: scenes where anchors appear on and scenes where no anchor appears. The results of the multimodal technique were significantly better when compared with the results from monomodal techniques applied separately. The tests were performed in four groups of Brazilian news programs obtained from two different television stations, containing five editions each, totaling twenty newscasts.

Keywords: scene segmentation, semantic segmentation, scene identification, scene boundary identification, multimodal technique.

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. 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 (Kodukula & Nazvia, 2011), aprendizado eletrônico (Zhang et al., 2006), bibliotecas digitais (Marchionini et al., 2006), videoconferência (Judge & Neustaedter, 2010), TV Digital (Souza Filho et al., 2007) e, por consequência, a TV Interativa (Athanasiadis & Mitropoulos, 2010), 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 a interação com o conteúdo multimídia da TV seja similar à 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 à 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 o uso de dispositivos com características distintas vinculadas a determinadas configurações de *hardware*, incluindo diminutos tamanhos de tela, restrições de processa-

mento e/ou memória, além de oscilações consideráveis na largura de banda e preferências adversas do usuário.

Alternativas que visam contornar essas questões e possibilitar ao usuário um acesso mais intuitivo e transparente ao conteúdo multimídia estão sendo desenvolvidas e pesquisadas, principalmente por uma área recente denominada personalização e adaptação de conteúdo (P&A), a qual tem potencializado o desenvolvimento de aplicações multimídia e, por consequência, contribuído com as áreas de Web e TV Digital Interativa (Magalhães & Pereira, 2004). Exemplos de adaptação ocorrem em sistemas que procuram decidir a versão de conteúdo ideal para apresentação e a melhor estratégia para gerar essa versão (LUM & LAU, 2002). A personalização, por sua vez, é vista como um caso particular de adaptação, pois os dados são adaptados para um único usuário. Em resumo, enquanto a adaptação tem como objetivo disponibilizar meios de acessar conteúdo multimídia a partir de diferentes condições de dispositivos, rede e ambiente computacional, a personalização estuda meios de customizar e/ou filtrar os dados segundo as preferências, interesses e necessidades de um usuário específico.

Nos últimos anos, o enfoque dos pesquisadores está centrado em personalização, oferecendo diferentes serviços que podem ser categorizados em seleção de conteúdo, sistemas de recomendação e sistemas de sumarização (Adomavicius & Tuzhilin, 2005). A seleção de conteúdo compõem os serviços que oferecem busca de itens multimídia a partir de critérios definidos pelos usuários, os quais podem ser baseados em anotações (tinta/caneta eletrônica, por exemplo) criadas durante o enriquecimento do conteúdo, atividade também vista como uma customização ou personalização de dados audiovisuais originais e manuais.

Na recomendação, a seleção dos itens de interesse é efetuada automaticamente, a partir de um perfil de preferências para cada indivíduo, geralmente representado por um documento textual que especifica dados pessoais e diferentes níveis de interesse do usuário para uma variedade de tópicos e assuntos relevantes. A extração das informações nos sistemas de recomendação pode ser realizada explicitamente, com intervenção do usuário, ou implicitamente pelo próprio sistema de maneira automática, baseando-se no histórico de interações do indivíduo.

Já na sumarização, o intuito é produzir uma versão modificada do conteúdo, com base em uma agregação de informações que são selecionadas a partir de critérios definidos pelo usuário e de seu perfil de interesses. Assim, a sumarização permite a criação de um sumário contendo apenas informações relevantes, seja com dados espaciais, por meio de objetos segmentados ou componentes intraquadro, ou temporais, como por exemplo, um noticiário contendo apenas notícias/cenas de determinadas categorias (*e.g.* saúde e educação).

Em geral, os sistemas de personalização apresentam uma necessidade em comum, que é o conhecimento das informações contidas no conteúdo. Também chamadas de metadados, essas informações têm a função de descrever a mídia em si, como formato, tipo de com-

pressão, tamanho de arquivo, além de disponibilizar dados informativos sobre o conteúdo sendo apresentado. Todavia, nesse último caso, as descrições podem variar em nível de granularidade ou detalhamento, assim como em nível semântico de representação (Snoek et al., 2005), dificultando a compreensão das informações e ocasionando, por consequência, um problema conhecido como lacuna semântica (Smeulders et al., 2000). Tal problema é caracterizado pela pouca ligação entre as informações de baixo nível obtidas com os metadados do conteúdo (histogramas, quantidade de movimento, localidade de uma cena, identificação de pessoas) e a interpretação do usuário para esse mesmo conteúdo.

Informações semânticas têm potencial para promover uma melhoria nos serviços de personalização, uma vez que podem formar um elo entre a representação computacional de um conteúdo e a interpretação dos dados por um determinado usuário, diminuindo, por sua vez, a lacuna semântica (Smeulders et al., 2000). Contudo, no caso dos vídeos digitais, para contornar a questão da lacuna semântica, é preciso anteriormente realizar a segmentação temporal do fluxo de vídeo.

Para obter informações semânticas, em geral, como uma etapa de pré-processamento, realiza-se a segmentação do vídeo (Hanjalic, 2004). Em particular, a segmentação semântica, também conhecida como segmentação de cenas, é ainda um tema de pesquisa em aberto, carecendo de investigação em técnicas e métodos que possam ser aplicados de modo a contribuir com a área de P&A.

Nesse contexto, pesquisadores têm voltado seus esforços para o uso de várias mídias, visando extrair o máximo possível de dados do fluxo de vídeo que possam servir de subsídios para a geração de informação semântica. No caso de identificação de estruturas de vídeo para o gênero telejornal (foco deste trabalho), a literatura reporta alguns trabalhos que exploram: a) reconhecimento de faces (Fang et al., 2006; Liu et al., 2009) e histogramas (Yu et al., 2007; Wang et al., 2008) para a identificação de âncoras (apresentadores de telejornal); b) análise do áudio associado que indique pausa e mudança de assunto/locutor (Wang et al., 2008; Liu et al., 2009); c) uso de informações textuais para detecção de mudança de cena com *closed-captions* (Ogawa et al., 2008), reconhecimento de fala (Wang et al., 2008); reconhecimento de caracteres (Fang et al., 2006)).

Todavia, essas técnicas simplificam a definição de cena em busca de melhores resultados, o que em alguns casos impede a correta segmentação de elementos que compõem a estrutura dos telejornais. Em geral, utiliza-se a aparição do âncora na imagem para determinar o início da cena. Contudo, cenas onde mais de um âncora aparece não são detectadas satisfatoriamente (Colace et al., 2005; Liu et al., 2009). Reconhecimento de faces tem sido usado para resolver esse problema (Zhao et al., 2006; De Santo et al., 2006b). Porém, a maioria das técnicas multimodais analisadas nos trabalhos relacionados não conseguem segmentar cenas corretamente em trechos de telejornal onde nenhum âncora aparece. Isso implica que tais técnicas não conseguem tratar adequadamente segmentos de vídeo com complexidade semântica elevada. Por exemplo, um trecho de telejornal

positionado entre duas aparições do âncora e contendo três segmentos, sendo que cada segmento trata de um assunto diferente. Tal trecho seria segmentado como uma única cena, quando o mais apropriado seria segmentar em três cenas diferentes.

Dessa maneira, o presente trabalho propõe uma técnica multimodal que segmenta cenas em telejornais empregando técnicas visuais, sonoras e textuais que, utilizadas em conjunto possam melhorar a semântica associada ao conteúdo. O principal objetivo da técnica é identificar as transições de cenas de acordo com um conceito mais geral, possibilitando com isso, contornar os problemas associados aos trabalhos relacionados e, consequentemente, contribuindo com a área de P&A.

Para a validação dessa abordagem foram utilizadas as medidas de avaliação precisão e revocação, já consolidadas pela área de análise de vídeo digital baseado em conteúdo, juntamente com uma base de telejornais compostas por quatro programas de telejornais distintos, cada qual com cinco edições, obtidos de duas emissoras de TV distintas, totalizando 20 telejornais brasileiros.

Os resultados obtidos demonstraram que a técnica desenvolvida consegue segmentar cenas mesmo em trechos de vídeo contendo segmentos com assuntos distintos, assim como em segmentos sem âncora. Além disso, os resultados confirmam que a técnica multimodal obtém melhores respostas do que as técnicas aplicadas isoladamente.

A principal limitação deste trabalho está relacionada ao desempenho da técnica, uma vez que este apresentou valores de precisão e revocação menores que os trabalhos relacionados. Contudo, uma ressalva importante deve ser feita: as definições de cenas são diferentes, sendo que as definições utilizadas neste trabalho abordam um conceito mais amplo. Isto torna difícil realizar uma comparação justa com as outras abordagens. Outro fator limitante que dificulta a comparação decorre do fato de não existir uma base de vídeos aberta e que contenha metadados associados ao conteúdo.

Como trabalhos futuros, a comparação de abordagens usando uma base de vídeos que contenha uma variedade maior de telejornais, considerando também os brasileiros, é uma alternativa para análise de resultados, desde que siga somente uma definição de cena. Ainda, investigar se esta técnica multimodal pode ser utilizadas em outras categorias, como documentários por exemplo, verificando também os esforços necessários para fazer tal mudança caso não seja possível utilizá-la de modo direto.

Este trabalho está organizado do seguinte modo. No capítulo 2, são apresentados os conceitos, definições e metodologias relacionados ao conteúdo do vídeo digital. O capítulo 3 descreve os trabalhos relacionados à segmentação de vídeo em cenas. No capítulo 4 é apresentado o sistema desenvolvido, as técnicas que o compõe e os resultados obtidos numa base de vídeo digital. Por fim, o Capítulo 5 recapitula o trabalho apresentado nesta dissertação ao sumarizar os resultados e contribuições alcançados, discutindo as limitações da abordagem proposta, e também sugestões para trabalhos futuros visando a continuidade deste trabalho.

Conceitos, Definições e Metodologias

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, foi 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 Youtube¹ 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 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 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.

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

¹<http://www.youtube.com>

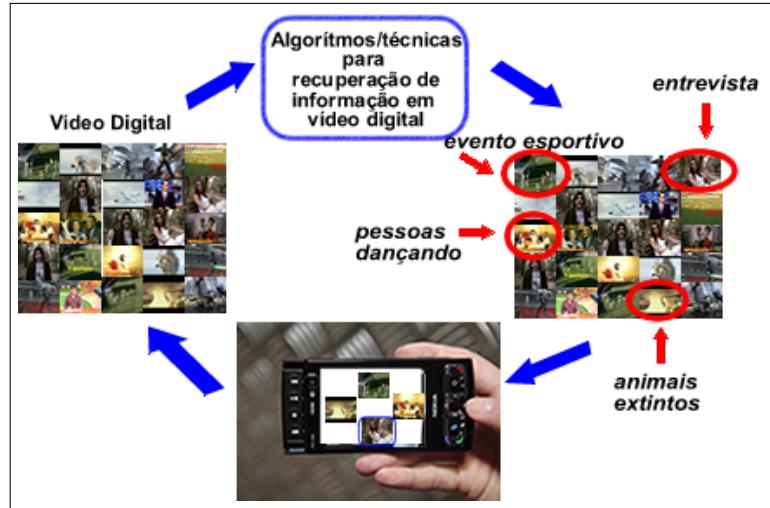


Figura 2.1: Algoritmos aplicados nos vídeos digitais que são desenvolvidos para facilitar a recuperação de seus conteúdos. Adaptado de Hanjalic (2004)

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 et al., 2001; Rui et al., 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 et al., 2001).

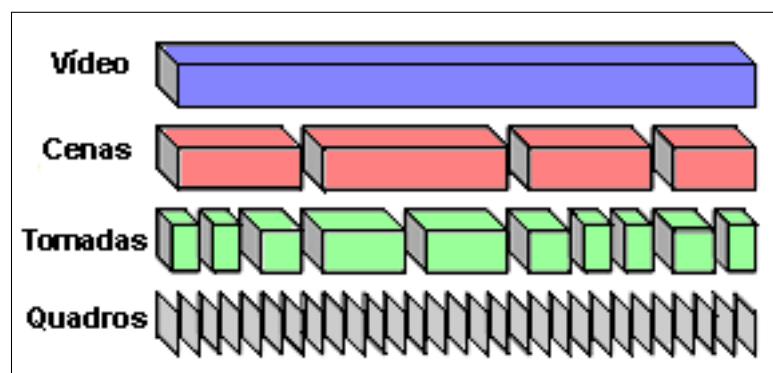


Figura 2.2: Estrutura do fluxo de vídeo digital

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, constituída por sequências de imagens (quadros) geradas por uma câmera do momento em que é iniciada a gravação até o momento do término da mesma (Hampapur et al., 1994). As cenas são definidas como um grupo de tomadas com conteúdo correlacionado

(Zhao et al., 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. telejornal, 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 et al., 2005; Oh et al., 2005; Zhao et al., 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 et al., 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.

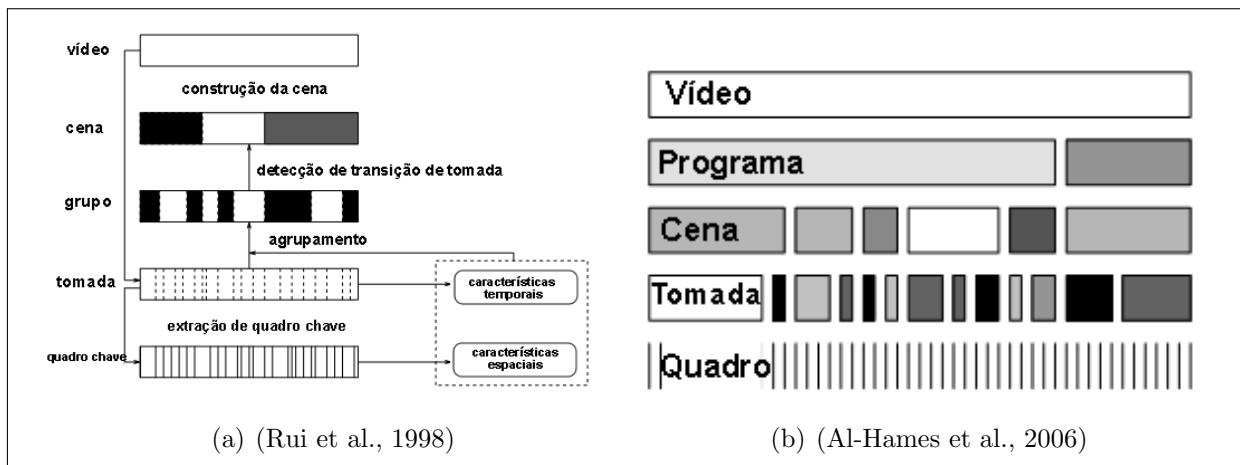


Figura 2.3: Outras representações para estrutura de vídeo digital

Apesar do processo de análise de vídeo ser realizado de modo qualitativo, 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 comprehensíveis dos segmentos de vídeo. O objetivo dessa etapa é comunicar de maneira eficiente e eficaz o conteúdo do vídeo para o usuário.

De modo resumido, 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 et al., 2005).

O método de segmentar um vídeo constitui da extração de suas partes relevantes. A segmentação pode ser classificada como espacial ou temporal (Magalhães & Pereira, 2004). A segmentação espacial foca as suas principais técnicas em dividir o vídeo considerando determinadas características espaciais (objetos), detecção de faces e reconhecimento de objetos que não estejam no plano de fundo. 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 deste 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 & 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 et al., 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 prática 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 & 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).



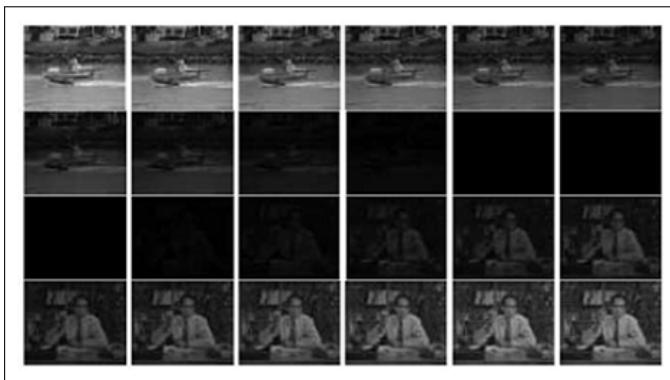
Figura 2.4: Transição abrupta de tomada

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 gradativo, isto é, estendendo em vários quadros (Figura 2.5).
- *Fade in* e *Fade out*: respectivamente, quando uma imagem escura clareia gradativamente ou o oposto, imagem clara escurece 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 et al., 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 2.5: Dissolução (Porter et al., 2003)

Figura 2.6: Transição *fade out* seguida por *fade in* (Koprinska & Carrato, 2001)Figura 2.7: Diferentes tipos de *wipes* (Joyce & Liu, 2006)

Devido a maior parte do processamento computacional ser dedicada a esses algoritmos 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 as 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 um vídeo é segmentado 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 a 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 Quadros-Chave

Um elemento importante na extração de informação utilizando características visuais é o quadro-chave (do inglês, *key-frame*). Segundo Yinzi et al. (2010), a extração de quadros-chave pode ser dividida em três classes distintas: baseada em amostragem, baseada em segmentos e baseada em tomadas. A abordagem baseada em amostragem é a mais simples e seleciona os quadros-chave capturando quaisquer quadros por meio de um intervalo fixo de tempo. A baseada em segmentos agrupa os quadros ou tomadas que possuem similaridades, de cor, textura e/ou movimento, em um conjunto de segmentos com um algoritmo de agrupamento, extraiendo quadros-chave de cada segmento. A complexidade do algoritmo e a dificuldade de determinar o número de segmentos são suas desvantagens. Por fim, a abordagem baseada em tomada é a mais comum, a qual realiza a segmentação temporal em tomadas e seleciona quadro(s)-chave para cada tomada aplicando técnicas que envolvem a seleção do primeiro quadro da tomada, média de *pixels*, média de histograma de cor ou considera o conteúdo. Tanto a abordagem de extração por amostragem quanto por tomada (no caso de extrair mais de um quadro-chave por tomada) podem ocasionar um volume muito grande de quadros-chave proporcionando redundância de informação e diminuindo a eficácia das técnicas do próximo nível da estrutura do vídeo, ou seja, técnicas de segmentação de cenas. A vantagem da segmentação baseada em tomada, usando um quadro-chave por tomada é diminuir a redundância e ser mais adequada a

tomadas com poucas mudanças ou movimentos de câmera, como acontece no telejornal (Zhang, 2002).

2.3.3 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 et al., 2008). Segundo o fluxo contrário dessa definição, nota-se que à compreensão semântica depende de estruturas denominadas cenas. Segundo Aner-Wolf & 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 & 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 et al., 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	<i>pitch</i> , espectro de frequência, características de sinais temporais, fones, <i>zero-crossing rate</i>
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 são 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).

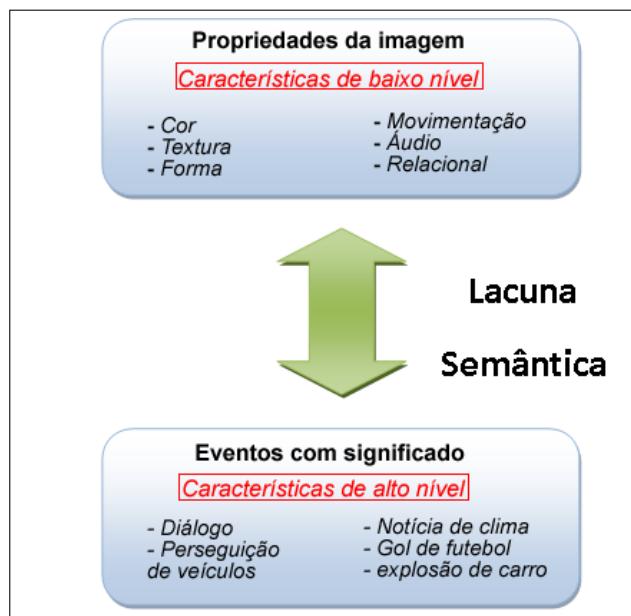


Figura 2.8: Representação da lacuna semântica. Adaptado de (Hanjalic, 2004)

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

As aplicações relacionadas a segmentação/identificação de cenas ocorrem em vários domínios, fornecendo diversos benefícios: em filmes menores a segmentação de cenas

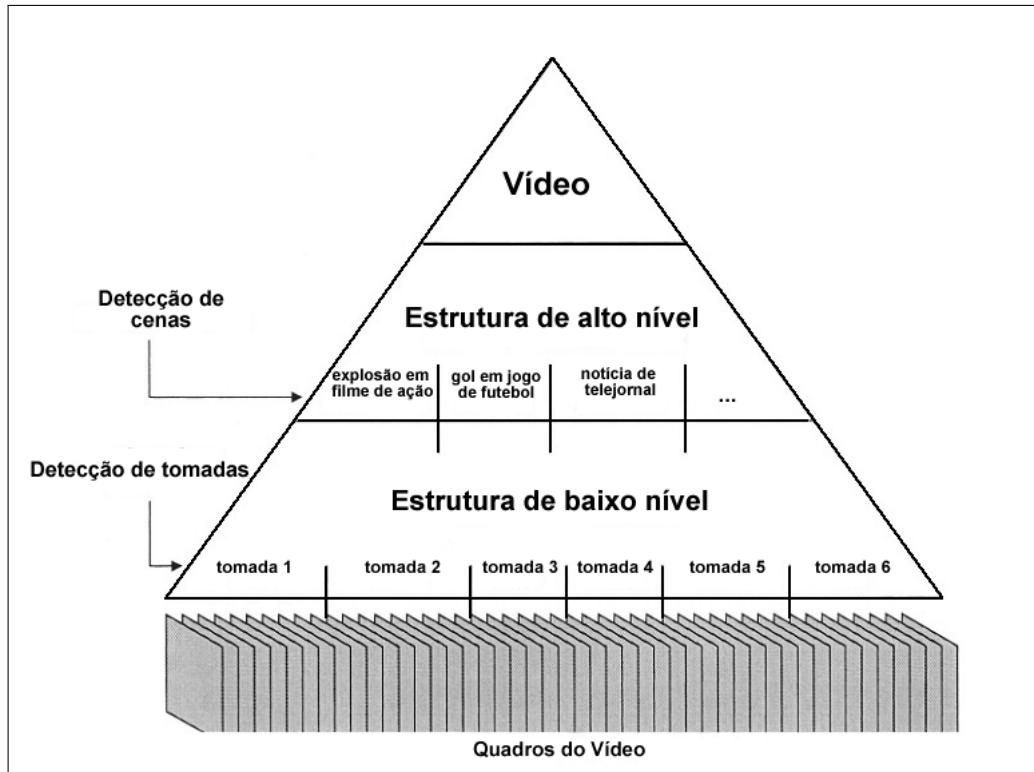


Figura 2.9: Pirâmide da estrutura do conteúdo do vídeo. Traduzido de (Hanjalic, 2004)

provê capítulos que correspondem a diferentes subtemas do filme; em vídeos de televisão, a segmentação pode ser usada para separar os comerciais dos programas comuns. Nos telejornais, a segmentação pode ser utilizada para identificar diferentes histórias jornalísticas (tal como clima, economia, política, 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).

Ao passo que as cenas possuem uma semântica associada ao seu conteúdo e esse conteúdo varia de acordo com os inúmeros domínios, os quais alguns foram relacionados anteriormente, Snoek & Worring (2005) propôs um modelo de granularidade chamado de indexação semântica hierárquica. Tal modelo separa os temas/domínios de vídeos já estudados na literatura acadêmica em 5 níveis, descritos a seguir:

1. **Propósito:** conjunto de vídeos que compartilham idéias similares;
2. **Gênero:** conjunto de vídeos que compartilham estilos similares;
3. **Sub-gênero:** um subconjunto de vídeos que possuem similaridades no conteúdo;
4. **Unidades Lógicas:** partes contínuas de um conteúdo de vídeo;
5. **Eventos nomeados:** pequenos segmentos que possuem um significado que não se altera durante o tempo;

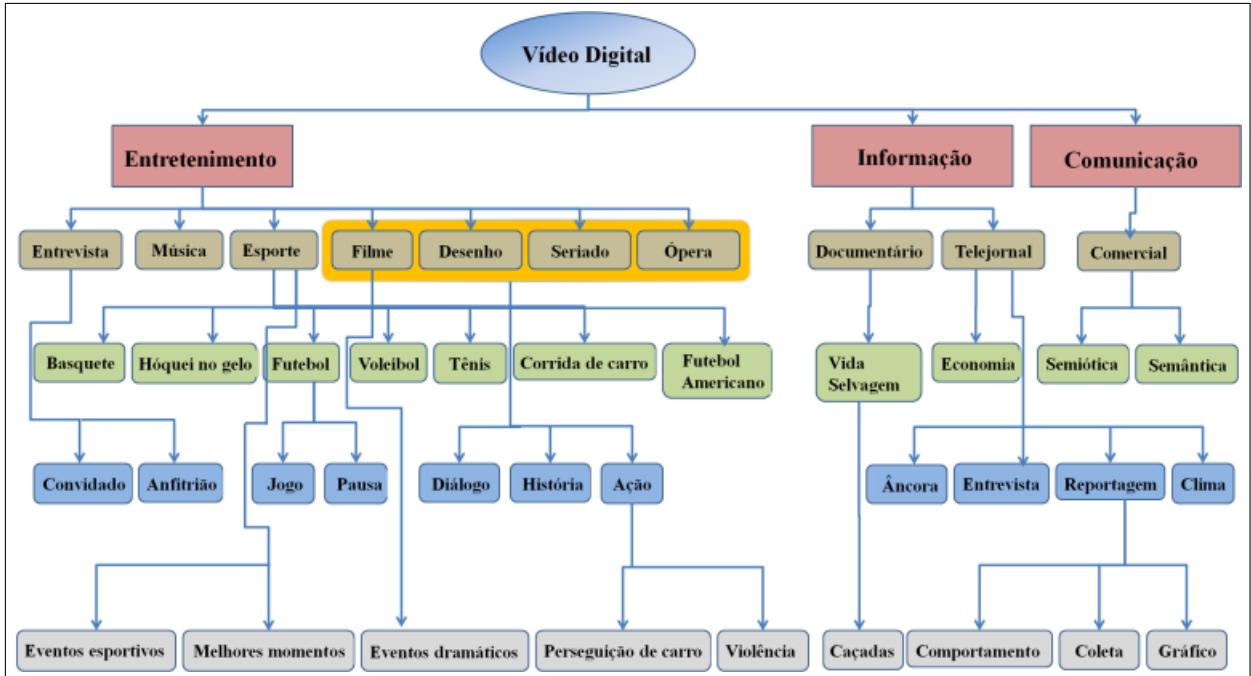


Figura 2.10: Modelo de indexação semântica hierárquica de vídeo proposto por Snoek & Worring (2005)

A Figura 2.10 apresenta os temas/domínios mais comuns no contexto acadêmico, assim como a disposição dos mesmos no modelo de indexação semântica hierárquica. No primeiro nível, o propósito do vídeo é composto por três seguimentos: entretenimento, informação e comunicação. Exemplos de gêneros (segundo nível) varia de filmes, telejornais à comerciais. O terceiro nível são compostos por diferentes sub-gêneros como filme de terror ou uma partida de hóquei no gelo. Exemplos de unidades lógicas, no quarto nível, podem ser diálogos em um filme de drama, o primeiro quarto em um jogo de basquete ou notícia de clima em um telejornal e, por último, no nível mais baixo, são exemplos de eventos nomeados variam de explosões em filmes de ação, gols em uma partida de futebol, a cotações de ações em notícias de economia de um telejornal.

2.3.4 Cenas em Telejornais

Como já mencionado, definir unidades correlacionadas semanticamente (cenas) não é uma tarefa trivial, pois existe o problema da lacuna semântica. Quando o gênero do vídeo analisado é o telejornal torna-se mais complicado porque este possui a peculiaridade de abordar assuntos bem distintos em seu conteúdo. Todavia, mesmo sendo um tema de pesquisa em aberto ainda hoje, trabalhos mais antigos já exploravam a importância de identificar esses segmentos. Altheide (1985) utilizou o termo “unidades de informação” para representar cenas, enquanto Graber (1990) utilizou o termo “cenas visuais” e, posteriormente, “unidades físicas” foi a denominação adotada por Riffe et al. (1998). O fator em comum nesses trabalhos foi a tentativa de descrever o significado destes segmentos uti-

lizando técnicas computacionais e/ou conceitos de alto nível. Entretanto, trabalhos mais recentes relatam que cenas em noticiários tanto é um conjunto de tomadas que retratam uma simples ação acontecendo em um mesmo local, quanto uma montagem que retrata um único conceito, tema ou idéia sem limitações de tempo e espaço (Choi & Lee, 2010).

Por consequência dessas definições, o conceito de cena empregado neste trabalho segue a mesma linha de Choi & Lee (2010) a qual representa um único tema ou idéia sem limitações de tempo ou espaço. Assim, cada notícia, vinheta ou comercial é uma cena diferente, pois entre o término de cada um destes segmentos e início do próximo ocorre uma mudança de tema/assunto, possivelmente alterando a correlação semântica e, portanto, ocasionando uma transição de cenas. Desse modo, considerando a estrutura dos telejornais (Figura 2.11) neste trabalho as transições de cenas acontecem quando ocorre:

- transição de notícias,
- transição de vinheta para notícia (ou vice-versa) ou
- transição de vinheta para comercial (ou vice-versa).

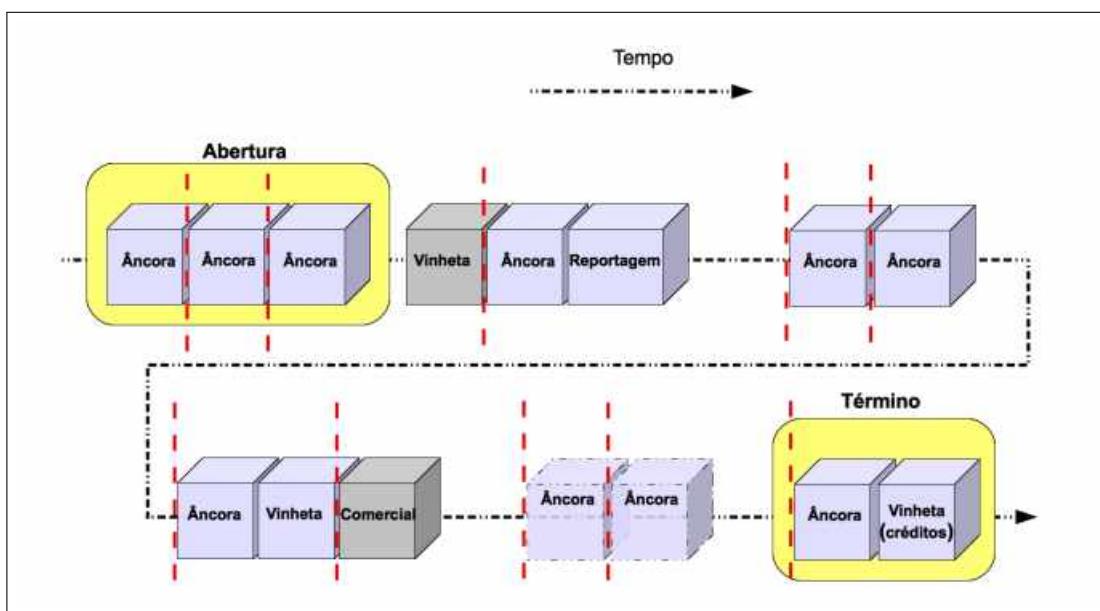


Figura 2.11: Composição temporal dos telejornais

De acordo com Chaisorn et al. (2003), a maioria dos telejornais possuem uma estrutura similar e bem definida, a qual geralmente começa com uma abertura composta de resumos das principais notícias abordadas. A principal parte do programa contém uma série de histórias organizadas por interesse geográfico (nacional ou internacional) e várias categorias como política, interesses sociais, finanças, esportes e entretenimento. Cada história pode ser vista como uma notícia que normalmente começa com a imagem do repórter âncora, sendo que ao longo do programa ocorrem períodos de vinhetas e comerciais publicitários entre blocos de notícias. Mesmo que a ordem das cenas seja um pouco

diferente de acordo com a transmissora/canal assistido, todos eles tem estrutura e categoria de notícias similares. Desse modo, a Figura 2.11 apresenta a composição/estruturas do telejornal em relação ao tempo, assim como os possíveis casos em que ocorre a transição de cenas (representadas por linhas tracejadas em vermelho na vertical). O detalhe nessa figura ocorre por conta do bloco transparente de âncoras na segunda linha de cima para baixo, pois representa as notícias em quem não ocorrem as imagens dos âncoras, somente a fala desse apresentador, transição também considerada no escopo deste trabalho.

No contexto relacionado à notícia, ocorre somente uma peculiaridade e acontece quando, por exemplo, há notícias consecutivas e derivadas de uma mesma categoria (economia, esporte, etc), pois podem ocasionar dificuldades em identificar seus respectivos inícios e o términos. Para exemplificar o conceito, imagina-se um cenário em que o âncora apresenta (em forma de gráficos) um aumento da inflação na economia brasileira. Na sequência, outro âncora aborda o tema da consequência da inflação para o consumidor, aparecendo imagens de um supermercado e no áudio de fundo um repórter relatando a alta de preços de determinados produtos. E, por último, o mesmo âncora apresenta um pequeno trecho de entrevista do Ministro da Fazenda explicando a causa da inflação. Mesmo que o tema seja economia e o contexto do cenário (idéia) esteja relacionado à inflação brasileira, a definição adotada neste trabalho considera que esse cenário é composto por três notícias, exatamente porque o assunto em questão (inflação) retrata ações distintas (gráfico estatístico, impacto no dia-a-dia do consumidor e a explicação de uma autoridade do Poder Executivo) e não possuem as suas respectivas sequências de tomadas em um mesmo local.

Uma observação importante está na abertura do telejornal e sua apresentação sucinta das notícias a serem abordadas posteriormente. Mesmo que a apresentação da notícia dure poucos segundos, esta é cumputada, obviamente, como cenas pelas mesmas razões citadas anteriormente. Outra observação está relacionada às vinhetas, pois elas atuam na separação de blocos do programa e podem ocorrer em três momentos: *i*) no início do programa, após a abertura do telejornal; *ii*) durante o programa como chamada para as próximas notícias (anteriormente ao intervalo comercial/propaganda publicitária); *iii*) ao final, quando são apresentados os créditos da edição, como o editor, núcleo de redação, chefe de produção, diretor de imagem, etc. No término, não há notícias, no entanto, como o âncora apresenta o final da edição do telejornal, considera-se que o fim do telejornal é uma cena. Os comerciais também fazem parte do telejornal e são caracterizados pela vinheta do fim de um bloco de notícias e início do próximo bloco. Logo, as notícias, vinhetas e comerciais são os três segmentos que compõem o gênero telejornal.

2.4 Métodos de Segmentação de Vídeo

Como apresentado na seção 2.3, a segmentação de vídeo temporal envolve a detecção de segmentos com menor duração de tempo e características semelhantes. Nesse contexto,

esta seção descreve os principais métodos utilizados na literatura para a segmentação de vídeo, os quais são também utilizados neste trabalho.

2.4.1 Histogramas de Cor

O histograma de cor de uma imagem é uma característica visual construída contando o número de pixels de cada cor, ou seja, um histograma de imagem é referente à densidade de probabilidade da intensidade da imagem. De acordo com Zachary et al. (2001) essa descrição é importante e nos permite usar métodos da teoria da informação para expandir a representação de imagens com base em seu conteúdo. Com isso, uma imagem discreta $I = F(N_1, N_2)$ de tamanho $N_1 \times N_2$ pode ser representada estatisticamente pela função de densidade de probabilidade:

$$p(i) \equiv p\{F(1, 1), F(1, 2), \dots, F(N_1, N_2)\} \quad (2.1)$$

Considerando que cada valor do pixel é independente estatisticamente de todos os outros valores dos pixels, então a função de densidade de probabilidade é fatorada como:

$$p(i) = p\{F(1, 1)\}p\{F(1, 2)\} \dots p\{F(N_1, N_2)\} \quad (2.2)$$

Para um conjunto discreto de valores, a interpretação de $p\{F(i, j)\}$ é desenvolvida na base de intervalo finito de possíveis valores para $F(i, j)$. Em uma imagem digital, estes valores são as possíveis cores de cada pixel, geralmente assumindo que a distribuição de cores de uma imagem segue uma distribuição uniforme, ou seja, cada cor tem probabilidade $1/M$ de ser atribuída a um pixel, onde M é a quantidade máxima de cores.

Cogitando a função de densidade de probabilidade fatorada, Yinzi et al. (2010) definiu o histograma de cor como:

$$h_{A,B,C}(a, b, c) = N \cdot Prob(A = a, B = b, C = c) \quad (2.3)$$

onde A , B e C são as três dimensões de cores (RGB, HSV,...), N é quantidade de pixels da imagem e $Prob$ é a probabilidade do pixel ter um valor de cor (a, b, c) .

A medida de similaridade escolhida para comparar os histogramas de cor das imagens do trabalho desenvolvido foi a Intersecção de Histogramas. Essa proposta foi apresentada por Swain & Ballard (1991) na indexação de cor com aplicação em reconhecimento de objetos. De acordo com Barla et al. (2003) esse método mede o grau de similaridade entre dois histogramas de cor, sendo adequado para processamento de imagens não escaláveis e na presença de objetos não segmentados, possibilitando a construção de eficazes sistemas

baseados em cor.

A seguir são descritas as duas técnicas de histogramas de cor baseado na intersecção de histogramas como medida de similaridade.

Histograma Global de Cor O histograma global de cor considera a intensidade de cor de todos os pixels da imagem. O espaço de cor utilizado foi o RGB, quantizado em 256 bits para cada canal de cor. A fórmula da intersecção de histogramas globais de cor e suas especificidades são descritas a seguir (Barla et al., 2003):

$$H_{(global)}(I, M) = \sum_{b=1}^3 \frac{\sum_{j=1}^n \min(I_j, M_j)}{\sum_{j=1}^n M_j} \quad (2.4)$$

onde $n=256$, I é o histograma de uma imagem modelo, no caso o âncora do telejornal, M é o histograma de um quadro-chave de uma tomada e b é o canal de cor no qual as duas imagens estão sendo comparadas ($b \in [R, G, B]$). O valor obtido da normalização fica no intervalo de 0 e 1. Portanto, quanto maior a similaridade entre as imagens, mais próximo o valor fica de 1.

Histograma Local de Cor O histograma local de cor, assim como o histograma global, calcula a intensidade das cores na imagem, entretanto, não considera os pixels em separados, mas sim um grupo deles, formando blocos ($B \times B$) de pixels. Neste trabalho, o tamanho do bloco foi definido em 16×16 pixels. Portanto, o cálculo de intersecção de histograma calcula os histogramas de cada bloco da imagem modelo e os compara com o bloco equivalente da imagem candidata, descrito como (Barla et al., 2003):

$$H_{(local)}(I, M) = \sum_{b=1}^3 \frac{\sum_{g=1}^G \min(I_g, M_g)}{\sum_{g=1}^G M_g} \quad (2.5)$$

onde G é a quantidade de blocos da imagem ($G = [(N_1 \times N_2)/B]^2$), I_g é o histograma local de cor no bloco g da imagem modelo no canal de cor b ($b \in [R, G, B]$), e M_g é o histograma local de cor da imagem candidata. Do mesmo modo, o valor obtido da normalização também fica no intervalo de 0 e 1.

2.4.2 Transformada Discreta de Wavelet

As transformadas de wavelets, também chamada de decomposição wavelets, podem ser vistas como mecanismos para decompor ou quebrar sinais nas suas partes constituintes, permitindo analisar os dados em diferentes domínios de frequências com a resolução de cada componente amarrada à sua escala. Resumidamente, pode-se dizer que na análise

wavelet, um sinal é decomposto nas funções derivadas da wavelet mãe em diversas escalas e deslocamentos temporais (Misiti et al., 1996). Dentre as principais wavelets mãe destacam-se: a wavelet Haar, a família Deubechies, Coiflets, Symlets, Morlet e Meyer.

Como a análise de Fourier, a representação wavelet fornece acesso a um conjunto de dados de vários níveis de detalhes, todavia, as wavelets diferenciam-se de Fourier no sentido que as diferentes frequências descritas pelas funções básicas da wavelet são locais ao invés de somente globais, como acontece com Fourier. Isso ocorre porquê essa técnica consegue distinguir as características locais de um sinal em diferentes escalas e, por translações, elas cobrem toda a região na qual o sinal é estudado. Por causa dessas propriedades únicas, as wavelets são usadas em análise numérica, reconhecimento de padrões, compressão de imagens e sons, computação gráfica, processamento de imagens, etc. Dentre as principais vantagens associadas ao uso de wavelets na área de processamento digital de imagens estão: as decomposições waveletes permitem uma boa aproximação da imagem original com poucos coeficientes; os coeficientes fornecem informação que é independente da resolução da imagem original, permitindo comparar facilmente imagens de resolução diferente; e decomposições rápidas e fáceis de computar, requerendo tempo linear no tamanho da imagem e pouco código (Wen et al., 1999).

As transformadas wavelets podem ser contínuas ou discretas. A transformada wavelet contínua (do inglês, *Continuous Wavelet Transform* - CWT), possui parâmetros de dilatação e translação que variam continuamente, ou seja, é aplicada a um sinal com resolução temporal infinita e, por conseguinte, precisa de infinitas escalas e deslocamentos temporais infinitamente suaves gerando assim infinitos coeficientes. A proposta da transformada wavelet discreta (do inglês, *Discrete Wavelet Transform* - DWT) é escolher um subconjunto de parâmetros de dilatação e translação que variam discretamente, baseadas em potência de dois. Computacionalmente, a DWT tem melhor eficiência (mais rápida e economiza memória), exatamente por ser composta por valores discretizados do sinal.

Portanto, na área de processamento digital de imagem é geralmente mais comum o uso da DWT para extrair características de textura e/ou reconhecimento de face, contudo, existem dois modos de decompor uma imagem bidimensional usando essa transformada: a decomposição padrão e a decomposição não-padrão. A decomposição padrão aplica a DWT unidimensional a cada linha de valores de pixels, resultando em um coeficiente de média e os coeficientes de detalhe para cada linha. Após, tratam-se estas linhas transformadas como se elas fossem uma imagem, e aplica-se a DWT unidimensional para cada coluna. Os valores resultantes são todos os coeficientes de detalhes, exceto por um único coeficiente que representa a média geral. Na decomposição não-padrão são realizadas operações de decomposição alternadas entre linhas e colunas. Primeiro aplica-se o cálculo da média nos pares horizontais e faz-se a diferença dos valores dos pixels em cada linha da matriz que representa a imagem. Depois, aplica-se o cálculo da média nos pares verticais e encontra-se a diferença para a coluna do resultado. Por fim, repete-se o processo

recursivamente apenas no quadrante contendo as médias em ambas as direções.

Outro modo de conseguir resultados mais eficazes é por intermédio da transformada wavelet rápida (do inglês, *Fast Wavelet Transform* - FWT), utilizando os coeficientes da DWT (Mallat, 1989). Também conhecida como codificador de subfaixa de canais, a FWT envolve a filtragem do sinal de entrada baseada na função wavelet mãe utilizada.

Começando com um sinal de entrada discreto, o primeiro estágio do algoritmo da FWT decompõe o sinal em dois conjuntos de coeficientes. Estes dois conjuntos são os coeficientes de aproximação, contendo informações de baixa frequência e os coeficientes de detalhes, contendo informações de alta frequência. O vetor dos coeficientes de aproximação é obtido através da convolução com o filtro passa-baixa e o vetor dos coeficientes de detalhes é obtido através da convolução com o filtro passa-alta. A operação de filtragem é seguida por uma dizimação diádica ou subamostragem por um fator 2, isto porque, após serem feitas as convoluções, o número de coeficientes é dobrado em relação ao sinal de entrada. Assim, o que operador de dizimação faz é eliminar todas as amostras de ordem ímpar dos vetores de aproximação e convolução, mantendo um número de coeficientes igual ao do vetor original.

Foi escolhido para este trabalho o uso da wavelet mais simples de Daubechies (daub4), gerada a partir de quatro coeficientes (Nievergelt, 1999):

$$(h_0, h_1, h_2, h_3) = \left(\frac{1 + \sqrt{3}}{4\sqrt{2}}, \frac{3 + \sqrt{3}}{4\sqrt{2}}, \frac{3 - \sqrt{3}}{4\sqrt{2}}, \frac{1 - \sqrt{3}}{4\sqrt{2}} \right) \quad (2.6)$$

A partir desses coeficientes constrói-se a função escala:

$$\phi(t) = \sqrt{2} \sum_{k=0}^{2N-1} h_k \phi(2t - k) \quad (2.7)$$

calcula-se g_n :

$$(g_0, g_1, g_2, g_3) = \left(\frac{1 - \sqrt{3}}{4\sqrt{2}}, \frac{-3 + \sqrt{3}}{4\sqrt{2}}, \frac{3 + \sqrt{3}}{4\sqrt{2}}, \frac{-1 - \sqrt{3}}{4\sqrt{2}} \right) \quad (2.8)$$

Assim, a wavelet de Daubechies é dada por:

$$\Psi(t) = \sqrt{2} \sum_{k=0}^{2N-1} g_k \phi(2t - k) \quad (2.9)$$

onde N é a quantidade de coeficientes (4), ϕ é a função escala e t é o tempo. Neste

trabalho, a wavelet de Daubechies foi calculada com o auxílio da biblioteca JWave² em Java.

A distância euclidiana foi a função de distância escolhida para medir a similaridade dos resultados das wavelets. Adotou-se essa função devido aos bons resultados obtidos durante a experimentação e, também, porque a literatura reporta experiências que sugerem sua boa adequação em aplicações de recuperação de informações (Zhang & Lu, 2003).

O cálculo para a distância euclidiana é dado pela seguinte fórmula:

$$D_{euclidiana}(I, M) = \sum_{b=1}^3 \sqrt{(Dwt_I - Dwt_M)^2} \quad (2.10)$$

onde Dwt_M é a transformada discreta de wavelet rápida aplicada na imagem modelo, Dwt_I é a transformada discreta de wavelet rápida aplicada na imagem que representa um quadro-chave do vídeo e b é o canal de cor na qual as imagens estão sendo comparadas ($b \in [R, G, B]$).

2.4.3 Root Mean Square

A maioria das características em nível de quadros de áudio são herdadas do tradicional processamento de áudio de voz/fala. Geralmente elas podem ser separadas em duas categorias: características no domínio do tempo, as quais são computadas diretamente das formas de onda, e da categoria de características no domínio da frequência, que são derivadas da transformada de Fourier das amostras nos quadros (Wang et al., 2000).

O volume, também referido como barulho/ruído/sonoridade (do inglês, *loudness*³), é a característica mais utilizada e com pouca complexidade computacional. Para exemplificar, o estudo de Liu et al. (1998a) descrevem uma técnica que extrai oito características de áudio, dentre elas o volume (com RMS), para classificação de programas de TV, que incluem jogos de futebol e basquete, comerciais, telejornais, etc. Ou ainda, Chen et al. (2003) criam técnicas, dentre elas uma taxa de silêncio por intermédio do volume (com RMS), para extraír cenas de ação e cenas com diálogos.

Visto como um indicador confiável de detecção de silêncio, o volume pode servir como um auxílio na segmentação de uma sequência de áudio e, normalmente, é aproximado pelo RMS da magnitude do sinal contido em cada quadro (Wang et al., 2000). Especificamente, a raiz da média dos quadrados (do inglês, *Root Mean Square-RMS*) revela a variação temporal da magnitude de um sinal em relação à distribuição do volume nos clipes de áudio. O cálculo é feito, como o próprio nome diz, por um conjunto de quadros do som

²<http://code.google.com/p/jwave/>

³Esse termo causa subjetividade, uma vez que é uma medida que depende da frequência da resposta do ouvido humano. Portanto, é mais correto utilizar o termo volume, tanto em inglês quanto em português.

e computando a raiz quadrada da soma dos quadrados dos valores das amostras desses quadros, representado por:

$$v(n) = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} s_n^2(i)} \quad (2.11)$$

onde N denota o tamanho do quadro, s_i denota a i -ésima amostra no n -ésimo quadro, com i e $n \in \mathbb{N}$.

2.5 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 Precisão (do inglês, *Precision*) e Revocação (do inglês, *Recall*). Não somente encontradas na segmentação de vídeo, essas medidas são comumente utilizadas na avaliação de desempenho dos algoritmos 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 (Fórmula 2.12 - Precisão) e também avaliando casos em que detectou-se transições onde não ocorreram (Fórmula 2.13 - Revocação). As medidas são descritas a seguir:

$$\text{precisão} = \frac{\text{num. verdadeiro positivos}}{\text{num. verdadeiro positivos} + \text{num. falso positivos}} \quad (2.12)$$

$$\text{revocação} = \frac{\text{num. verdadeiro positivos}}{\text{num. verdadeiro positivos} + \text{num. falso negativos}}, \text{ onde} \quad (2.13)$$

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 & Ribeiro-Neto (1999) relatam que uma estimativa apropriada para a revocação 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 resol-

vido com uma abordagem que combine essas duas medidas (e.g. O significado Harmônico⁴ e a Medida E⁵).

Zutschi et al. (2005) também descreve algumas limitações sobre o uso dessa avaliação ao relatar que a utilidade prática dessas medidas estão sendo questionadas sob o ponto de vista 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, a revocação 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 da precisão e revocação 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).

2.6 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 relacionados 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, além de existir várias categorias semânticas de vídeos.

Definições mais específicas para o gênero telejornais foram apresentadas com o intuito de exemplificar e explicar os conceitos deste trabalho, assim como também ocorre com os métodos de segmentação de vídeo. Para avaliação de técnicas que realizam segmentação

⁴Composta 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.

⁵Por meio de uma constante adicionada à formula, o usuário pode definir qual das duas medidas é mais importante, funcionando como um peso.

semântica em vídeos, frequentemente são utilizadas medidas qualitativas como a precisão e a revocação, as quais tem o intuito de verificar a eficiência da metodologia empregada.

Segmentação de Cenas em Vídeo Digital

3.1 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, hipermídia¹, processamento de linguagem natural, entre outras, demandam grandes esforços para diminuir a lacuna entre a capacidade de interpretação humana e 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 encontradas em temas relacionados a segmentação de cenas. Essas características fornecidas pelo vídeo permitem diversas classificações baseado em seu conteúdo e podem ser representadas pelos gêneros de vídeos (esporte, filme, telejornal, documentário, etc.) ou ainda, caso exista, o tipo de tecnologia abordada (*e.g.* padrões MPEG, H.26x, etc.). Contudo, a classificação utilizada neste 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 trabalhos usando a classificação mencionada analisando, principalmente, vídeos do gênero telejornais (foco do trabalho), incluindo o estado da arte das técnicas multimodais. É importante enfatizar que os termos **identificação** e **segmentação** estão relacionados à identificação da transição de cenas, ou seja, identifica-se em qual momento do vídeo uma cena termina e a outra começa.

¹Hipermídia 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).

3.2 Características Visuais

Dentre as áreas responsáveis pela recuperação por conteúdo, a Recuperação de Conteúdo Baseada em Imagem (do inglês, *Content Based Image Retrieval-CBIR*) é considerada 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 maneiras distintas, ocasionando uma subjetividade que torna difícil o trabalho de extração de informação realizado pelo computador.

Zeng et al. (2010) identificou aspectos similares nos trabalhos que realizam segmentação de cenas usando características das imagens dos vídeos e classificou-os em três categorias:

- Abordagem baseada em agrupamento de tomadas: considera que uma cena é formada por um conjunto de tomadas similares semanticamente, os trabalhos que fazem uso desse método utilizam essa similaridade entre as tomadas para fornecer algum vestígio ou indício de agrupamento (Cao, 2007; Zhu & Liu, 2009; Dunlop, 2010; Zeng et al., 2010).
- Abordagem baseada em detecção de transição: nesta abordagem, transições entre tomadas são consideradas como candidatas a transição de cenas (uma vez que transição de cena é uma transição de tomada, mas nem sempre uma transição de tomada é uma transição de cena) e transições falsas são removidas checando a coerência da semelhança entre tomadas diferentes (Gu et al., 2007; Zhu & Liu, 2008b; Chen & Li, 2010; Huang & Zhang, 2010).
- Abordagem baseada em modelo: esta abordagem possui a idéia de que para agrupar N tomadas em K cenas é equivalente a estimar parâmetros de um determinado modelo (Tan & Lu, 2002; Zhai & Shah, 2005; Ren et al., 2010).

Relacionado ao agrupamento de tomadas, técnicas de textura de CBIR como a Transformada Discreta de Wavelet foi empregada por Zhu & Liu (2009) especificamente na etapa de seleção dos quadros-chave. Em conjunto com essa técnica, outra técnica de cor (histograma de variação de nível de cinza) foi usada para que ambas, e em conjunto com variáveis temporais, pudessem agrupar as tomadas do vídeo. A sua principal vantagem está em detectar mais cenas do que as existentes no vídeo². Outros trabalhos (Cao2007, Dunlop2010) utilizam classificadores binários baseados em SVM (do inglês, *Support Vector Machine*) para fazer agrupamentos. Cao (2007) extrai a cor e a textura dos quadros-chave para esses classificadores, agrupando tomadas de um documentário em cenas usando diferentes classes semânticas. Como resultado foi feito uma comparação com outras técnicas, obtendo melhores resultados, com a ressalva que pode ocorrer resultados errôneos caso haja cenas adjacentes com o mesmo conteúdo semântico. Esses

²evento denominado de *over segmentation*, do inglês.

mesmos classificadores também criaram classes semânticas no trabalho de Dunlop (2010), porém rótulos foram adicionados a determinados tipos de quadros-chave (tipo externo), descrevendo os componentes da cena de acordo com uma piramide espacial. A pequena quantidade de classes semânticas e a não análise de quadros-chave do tipo interno foram os limitantes desse trabalho. Zeng et al. (2010) realizaram o agrupamento de tomadas inserindo autocorrelograma de cor (HSV), distância entre quadros consecutivos e também variáveis temporais em uma matriz de similaridade. Mesmo obtendo um número ínfimo de cenas não detectadas detectou-se muitas cenas erroneamente, com exceção de telejornais, os quais obtiveram bons resultados.

Como representantes do segundo grupo, Zhu & Liu (2008b) detectaram e removeiram os quadros-chave que não possuiam informação útil por intermédio da técnica de comparação com um modelo (do inglês, *template matching*), detectando cenas a partir da similaridade visual e temporal das tomadas que não tiveram seus quadros-chave excluídos. Os níveis de precisão e revocação ficaram acima de 80% tendo como base de dados um vídeo de entrevistas e quatro filmes de ação. Na mesma linha, Gu et al. (2007) elaboraram um procedimento que identifica as transições com um nível mais alto de similaridade, descartando as restantes, com o uso da técnica de Minimização de Energia baseada em Segmentação (do inglês, *Energy Minimization Based Segmentation*), revelando melhores resultados em gêneros de filmes do que em vídeos caseiros. Chen & Li (2010) fizeram uso de características de cor e temporais das tomadas para realizar os agrupamentos, fazendo com que as características de intensidade de movimentação das tomadas excluam as transições de cenas redundantes. Um ponto negativo desse trabalho foi a proposta de avaliação, de somente dois filmes, sendo um pequeno segmento de cada um. Huang & Zhang (2010) calcularam a similaridade entre os quadros-chave de tomadas anteriores e posteriores a um determinado quadro-chave, a qual é, por sua vez, uma medida de exclusão ou inclusão para identificar a ocorrência de uma transição cenas. Como a técnica também faz detecção de tomadas, se acontece erro nessa etapa, o erro é replicado na segmentação das cenas. A eficácia de transições de cenas também diminui quando coincide com uma transição gradual de tomadas, pois essa última é um gargalo da técnica.

Alguns trabalhos utilizam determinados algoritmos que geram modelos a fim de identificar as cenas. Ren et al. (2010) geraram um modelo de cena a partir da técnica do modelo de surgimento de superpixels com características de imagens de baixo nível, notadamente em vídeos com cenas urbanas. Como resultado a eficácia de cenas inferidas por esse modelo é melhor do que as nomeadas manualmente em classes semânticas. A técnica de Cadeia de Markov de Monte Carlo (do inglês, *Markov Chain Monte Carlo*- MCMC) auxilia no processo de modelagem, como acontece com Zhai & Shah (2005) ao formular a segmentação de cena como um problema de inferência Bayesiana. Mesmo utilizando o MCMC para solucionar o problema, a técnica fica restrita a quantidade de tomadas, pois quanto menor o número, menos eficiente é a técnica. Tan & Lu (2002) também consi-

deraram o MCMC e aglomera as cenas com o auxílio Modelo de Mistura Gausiana (do inglês, *Gaussian Mixture Model*). Nesse trabalho, cada cena é modelada com uma densidade Gausiana, levando em conta que características visuais similares pertençam a uma mesma cena. Assim, ficou comprovado que essa abordagem consegue descobrir semântica em vídeos de esportes, no entanto, para outros gêneros como vídeos caseiros ou filmes, apenas características de tomadas individuais não são suficientes.

No gênero de telejornais, a detecção de âncora(s) é uma abordagem muito utilizada para indicar transições de cenas, com a maioria dos trabalhos se baseando no pressuposto que diferentes imagens de âncoras compartilham o mesmo plano de fundo. Trabalhos iniciais, realizados por Zhang et al. (1994), construíram três modelos de âncoras para tomadas de âncora: tomada, quadro e região. A tomada de âncora é modelada como uma sequência de modelos de quadros e um quadro é modelado como um arranjo espacial de regiões. Como resultado, os autores perceberam que os modelos variam de acordo com o canal/transmissora de TV, sendo difícil construir todos os possíveis modelos para todos os diferentes telejornais. Posteriormente, fei MA et al. (2001) propuseram, um método baseado em detecção de borda para localizar as tomadas dos âncoras, o qual utilizou o operador DoG e generalizou a transformada de Hough (GHT) para equiparar o contorno dos âncoras. A desvantagem dessa abordagem é que consome muito tempo.

Embora histograma de cor seja uma técnica de extração de informação em imagens usada para identificar qualquer tipo de tomada de âncora (Lee et al., 2011), outras pesquisas relatam detecção de âncoras para um tipo específico de padrões de imagens que contenham planos de fundo dinâmico com a figura do âncora. Divididas em duas categorias, plano de fundo dinâmico parcial (Figura 3.1(a) e 3.1(b)) e plano de fundo dinâmico global (Figura 3.1(c) e 3.1(d)), Zheng et al. (2009) criaram uma técnica para detecção de ambos os casos utilizando um algoritmo que divide as imagens dos âncoras em sub-blocos, calculando seus respectivos histogramas com os histogramas equivalentes dos modelos de imagens de âncoras existentes, afim de identificar similaridades espaciais entre as imagens. Como resultado, o algoritmo mostrou bom desempenho para quaisquer categorias de plano de fundo dinâmico em telejornais japoneses.

Outra técnica muito usada para detecção de tomadas de âncoras é o reconhecimento de face (Lan et al., 2004; De Santo et al., 2006a; D'Anna et al., 2007). Essas abordagens possuem uma boa taxa de detecção, mas não são as melhores escolhas devido à sua inherente complexidade de algoritmos detectores de face. Todavia, há estudos que extraem características de textura com wavelets (yu Chen et al., 2010) para posterior reconhecimento de face, demonstram melhores resultados na identificação da figura do âncora.

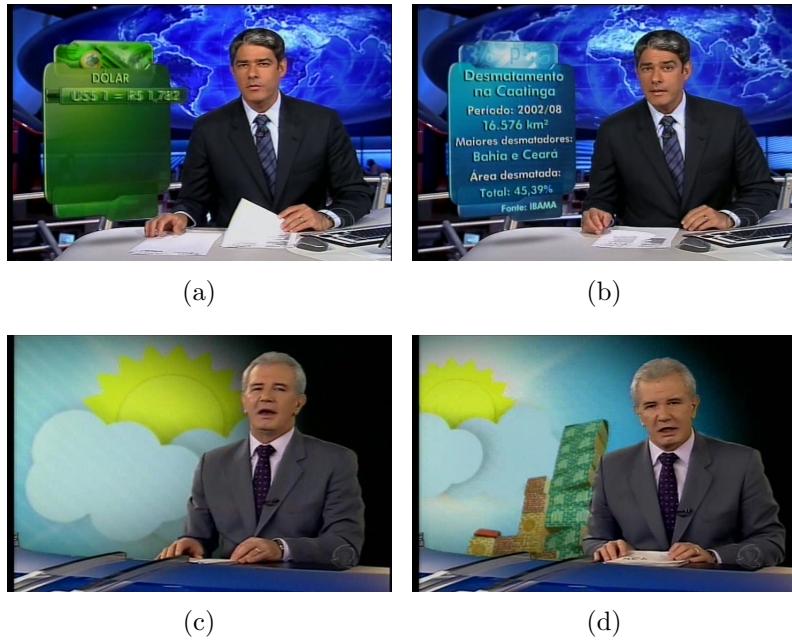


Figura 3.1: Âncoras com planos de fundo dinâmicos

3.3 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 segmentação em cenas somente com áudio é possível, apesar da forma de fazê-lo não ser tão intuitiva se comparada às mídias visual e de 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 & Chen (2006) é possível detectar mudanças de cenas em filmes mesmo sem acesso ao conteúdo visual do vídeo, apenas com o som, possibilitando estruturar o vídeo, mesmo que a fala (linguagem) não seja a da mesma 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 et al. (2000) relatam 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 3.2).

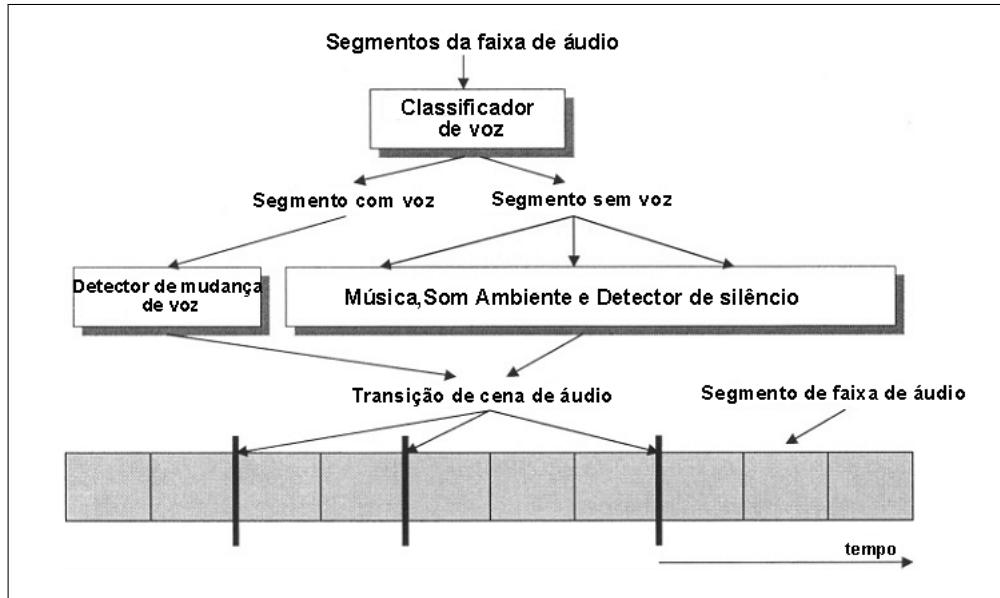


Figura 3.2: Visão geral dos métodos de detecção de cena com áudio. Traduzido de Jiang et al. (2000)

Pode-se dizer que os trabalhos mais recentes envolvendo detecção de cenas somente com áudio realizam identificação de cenas quando conseguem classificar o som em eventos coerentes com o conteúdo (gênero) do vídeo avaliado. Lu et al. (2006) adotaram 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 investigou segmentos de áudio, de programas de TV, os quais ocasionam transições de cenas. Entretanto, 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 et al. (2005). Um vetor de características armazena esses segmentos em grupos, transformando-os em índices. As cenas com seus respectivos índices comporta-se como segmento modelo para serem comparadas com amostras de som do vídeo original. A avaliação do resultado foi feita em dois gêneros, entretenimento e telejornais. 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 & Chen (2006) também fizeram uso da classificação do 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.

O aumento na demanda por dados comprimidos foi uma razão para que Yu et al. (2009) adotassem o formato MP3 para identificar a transição de cenas, usando para isso uma taxa baixa de ruído SNR³. Juntamente com a matriz de Transformada Discreta de Cosseno Modificada (do inglês, *Modified Discrete Cosine Transform* - (MDCT)) e um conjunto de técnicas de extração de dados em domínio comprimido de dados, foi implementada uma técnica que detecta segmentos de ruído puro, fala e música, mesmo quando o SNR é tão baixo quanto 0dB (decibéis), resultado comparável aos obtidos em domínios sem compressão.

Por fim, Zhang & Wang (2010) determinaram ocorrência de cenas de áudio para transições de programa de TV baseado em algoritmo para detecção de comerciais de TV. Duas técnicas foram aplicadas neste estudo, detecção de silêncio e um algoritmo de espalhamento de áudio. Resultados com precisão acima de 80% e revocação acima de 95% foram obtidos, além de proporcionar uma complexidade computacional pequena, demorando apenas 433 segundos para segmentar 9 horas de vídeo.

3.4 Características Multimodais

3.4.1 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 et al., 2000).

Com o benefício de incluir o som na recuperação de arquivos multimídia, Shao et al. (2006) adotaram 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, por intermédio da análise de conteúdo visual (histograma de cor), alinhando-os posteriormente. A avaliação foi centrada no usuário e obteve resultados comparáveis a sumarização manual. Na pesquisa de Dong & Li (2006) fizeram uso das mesmas técnicas no fluxo de áudio, como o *zero-crossing rate* (ZCR), que os autores ante-

³Signal-to-noise ratio

riores, com o objetivo de detectar cenas em documentários por meio de intervalos sonoros. Um desafio nesse trabalho é quando ocorre uma mudança abrupta no áudio, pois o algoritmo pode, erroneamente, detectar a voz do narrador como o som ambiente. O estudo de Coimbra & Goularte (2009) faz uso de uma técnica de característica visual (histograma global de cor) e uma de áudio (detecção de silêncio) para segmentar telejornais. Os resultados apresentados compararam o resultado das segmentações realizadas em separado e os resultados com um algoritmo que faz a união das duas técnicas por meio de constantes de tempo (Figura 3.3). Embora, a revocação tenha ficado maior, o algoritmo de união conseguiu identificar precisamente mais de 80% das cenas em todos os telejornais.

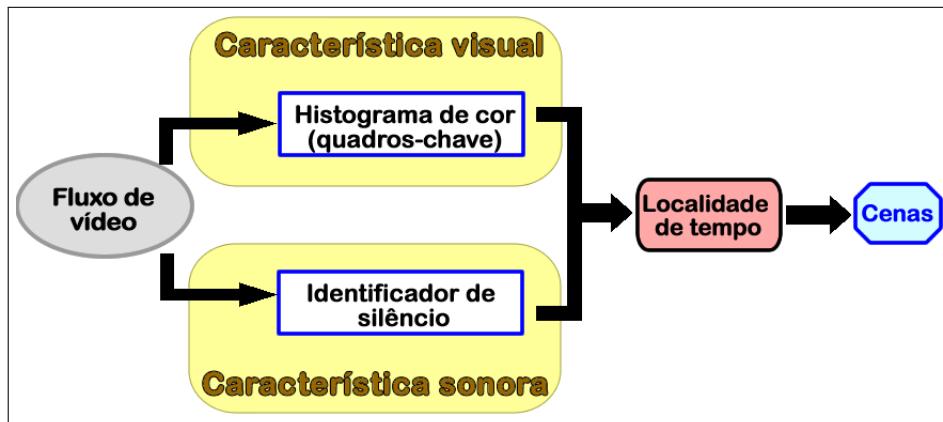


Figura 3.3: Arquitetura da técnica de união das características visuais e de áudio (Coimbra & Goularte, 2009)

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 de entrevistas e piores para telejornais.

3.4.2 Características Visuais e Textuais

Alguns trabalhos visam a identificação de cenas explorando somente as características visuais e textuais do fluxo de vídeo. De fato, Misra et al. (2010) relatam que os trabalhos

melhores avaliados, utilizando a TRECVID como base de dados, fizeram uso de ambas as características, visual e textual. Do mesmo modo, Chua et al. (2004) citam que as técnicas de áudio, em comparação com as técnicas de identificação de face e movimento (ambas características visuais), representam os maiores índices de erros para essa mesma base.

Visto que já foram mencionados os conceitos relacionados à extração de características visuais na seção 3.2, faz-se necessário o mesmo para as características textuais. Assim, a Tabela 3.1 apresenta os principais meios de se obter informação textuais, bem como as vantagens e desvantagens de cada um.

Tabela 3.1: Tabela das principais características textuais com suas respectivas vantagens e desvantagens (adaptado de Brezeale & Cook (2008))

Características Textuais	Vantagens / Desvantagens
<i>Closed-captions</i>	Alta eficácia quando não produzido em tempo real, alta dimensionalidade, extração computacionalmente barata
Reconhecimento de fala (ASR)	Alta taxa de erros
Reconhecimento de caracteres (OCR)	Possibilita a extração de texto em trechos de vídeos que não ocorrem diálogos, computacionalmente caro

Juntamente com o uso de características visuais, a maioria dos trabalhos abordam a técnica de reconhecimento de fala (do inglês, *Automatic Speech Recognition*- ASR) Koskela et al. (2009), em grande parte devido ao uso da base TRECVID, pois esta disponibiliza este tipo de metadado, em forma de arquivo de texto associado à cada vídeo. Por conseguinte, os restantes dos trabalhos exploram o reconhecimento de caracteres (do inglês, *Optical Character Recognition*-OCR) (Yu et al., 2007) e *closed-captions* (Ogawa et al., 2008).

Após a análise das oficinas dos anos de 2006, 2007 e 2008 do TRECVID, Koskela et al. (2009) indicam que o uso de conceitos semânticos associados a informações textuais obtidas por reconhecimento de fala (ASR) fornecem resultados melhores quando comparados com a combinação de recuperação de vídeo baseado em conteúdo e ASR. Um aspecto interessante desse trabalho é o desempenho ruim da recuperação baseada em texto quando analisada sozinha, fato devido ao reconhecimento errôneo de palavras. Outra comparação de abordagens foi realizada por Misra et al. (2010), os quais compararam os resultados de identificação de âncoras utilizando características visuais de cor do MPEG-7 e informações textuais obtidas por *closed-captions*. A integração ocorre quando

as transições de ambas as abordagens estão em uma janela de tempo de um segundo de distância entre cada uma, formando apenas uma transição. Essa integração consegue melhores resultados para ambos os conjuntos de telejornais americanos CNN e ABC.

No trabalho de Yu et al. (2007) histogramas de cor e textura de Gabor são utilizados para detectar âncoras e compõem o módulo de características visuais, enquanto o módulo textual é formado por ASR e OCR. A informação multimodal é composta pela união de ambos os módulos seguindo uma abordagem de *ranking* com pesos específicos para cada um. Os resultados foram avaliados na conferência TRECVID de 2005 e 2006, obtendo o quarto melhor resultado dentre os trabalhos que fizeram busca manual em 2005, e o sétimo em 2006. Ogawa et al. (2008) também faz uso de histogramas de cor para identificar cenas similares em telejornais de países distintos. As similaridades entre as palavras-chaves dos *closed-captions* constituem a informação textual, sendo a integração das abordagens também baseada em uma soma das suas respectivas similaridades. Assim como os trabalhos anteriores, foram comparadas as técnicas aplicadas em separadas e juntas, com melhores resultados para a união de ambas, mas ainda com falsos positivos quando ocorre vinhetas.

Hoi & Lyu (2007) propuseram um arcabouço multimodal baseado em *ranking*. Na parte visual foram utilizadas características de cor (momento de cor), forma (histograma de borda) e textura (transformada wavelet) e na parte textual, um analisador com uma lista de palavras de parada é aplicado no texto fornecido por ASR. Para a construção do ranking multimodal foram combinadas abordagens visuais e textuais juntamente com um método de aprendizado supervisionado (SVM), obtendo melhora de 40% com uso do *ranking* comparado com somente a técnica textual. Wavelets de textura e momentos de cor também são utilizadas por Xie et al. (2007). Um sistema baseado na frequência de palavras obtidas por ASR é integrado às características visuais, proporcionando resultados melhores que os propostos para busca de tópicos no TRECVID de 2005 e 2006.

3.4.3 Características Audiovisuais com Texto

Em oposição às técnicas de extração de informação em texto dos trabalhos de características visuais e textuais (ASR), a maioria dos estudos que incluem as três mídias do vídeo utilizam métodos baseado em OCR (Liu et al., 2009; Hua-Yong & Tingting, 2009; Jianping et al., 2009). Assim como os trabalhos que usam mais de uma mídia, as metodologias são similares, sejam nas abordagens visuais com identificação da imagem do âncora com reconhecimento de face (Zhao et al., 2006; Jianping et al., 2009) e histogramas (Zhao et al., 2006; Hua-Yong & Tingting, 2009), nas técnicas de áudio com detecção de silêncio (Zhao et al., 2006; Jianping et al., 2009; Hua-Yong & Tingting, 2009) e identificação ou mudança do locutor (Colace et al., 2005; Zhao et al., 2006), além de técnicas de integração das mídias com aprendizado de máquina (Colace et al., 2005; Jianping et al.,

2009) ou abordagens com *ranking* (Zhao et al., 2006; Wang et al., 2008). Apesar de apresentar uma metodologia detalhada, o trabalho falha em não apresentar maneiras de validação dos resultados obtidos, atendo-se apenas à apresentação das funcionalidades do arcabouço.

Hua-Yong & Tingting (2009) compararam os resultados das técnicas de mídia aplicadas separadas com a técnica multimodal e conseguiu uma melhora de cerca de 11% na precisão e 5% na revocação usando a multimodalidade. Os estudos foram baseados em técnicas de OCR para informação textual, comparação de histogramas para visual e detecção de cliques de silêncio com técnicas de extração de energia e *zero crossing rate* (ZCR). Mesmo obtendo resultados satisfatórios, a base de vídeo é restrita a apenas três telejornais de uma mesma emissora, além da fala de entrevistados causarem muitos falsos positivos com o OCR. A técnica de comparação de histogramas de cor para identificação da figura do âncora também fez parte do trabalho de Liu et al. (2007a), assim como OCR para texto e detecção de silêncio entre as notícias.

Fazendo uso de uma base de dados mais extensa, ao contrário da proposta anterior, estudos realizam os testes e validação de suas técnicas em base de vídeos com mais de uma emissora (Zhao et al., 2006) e também telejornais de países diferentes (Jianping et al., 2009), gerando técnicas mais abrangentes. Zhao et al. (2006) abordaram o uso de características de texturas e cor, dentre elas histograma local de cor, para reconhecimento do âncora e algoritmos para reconhecimento de face como técnicas visuais. Identificar os locutores e procurar momentos de silêncio auxiliaram a técnica multimodal, assim como abordagens para ASR e OCR para informação textual. Duas técnicas de integração de mídias foram elaboradas, uma com pontuação considerando aspectos das características em separado e outra de *ranking* agrupando pesos em uma única lista. A técnica de *ranking* obteve melhores resultados quando analisados em uma base de 60 horas de telejornais das emissoras CNN e ABC, mesmo com o reconhecimento de fala apresentando problemas, não sendo fidedigno à fala do locutor. Com uma base de telejornais de 15 horas do E.U.A e da China, Jianping et al. (2009) usaram reconhecimento de face, classificação de áudio com momentos de silêncio, OCR, intensidade de movimento e, por fim, classificação bayesiana para integrar todos os atributos. De modo geral, essa técnica obteve desempenho melhor que as outras duas abordagens comparadas no trabalho, contudo algumas notícias apresentadas sem pausa pelo âncora não foram detectadas e detectadas erroneamente algumas notícias com dois âncoras.

Liu et al. (2009) consideraram a legenda das imagens (OCR) a parte principal do sistema multimodal, mesmo considerando que perto das transições tenha momentos de silêncio e/ou mudança de locutor e que a figura da imagem do âncora, identificada por técnicas de reconhecimento de face, apareça na maioria do início das notícias. A técnica multimodal detecta cenas caso duas técnicas indiquem que em um determinado momento ocorre transição, com exceção da técnica de texto, que é capaz de identificar sozinha

essa transição. Foram analisados as taxas de erros de segmentação das cenas e a técnica multimodal apresentou o melhor resultado com a menor taxa. Mudança de locutor no áudio é o método também utilizado por Colace et al. (2005) como característica de áudio. Histogramas de cor global para detectar mudanças no plano de fundo e ASR para extrair informação textual completam as características que foram utilizadas para obter maior carga semântica dos vídeos. HMM foi a técnica adotada para integrá-las, formando a técnica multimodal. Uma desvantagem desse estudo ocorre na definição das cenas do tipo notícia, a qual é descrita como sempre tendo a imagem de um âncora no início, a qual é muito restrita, mesmo para a base de oito telejornais italianos analisados.

A análise não somente de telejornais mas de programas de TV em geral foi efetuada por Wang et al. (2008) em cinco noticiários de emissoras diferentes, americanas e chinesas. As técnicas visuais são restritas a histogramas de cor e borda globais e locais classificadas com SVM, silêncio, ZCR, Pitch e outras características fazem parte das características de áudio e o texto foi obtido por ASR e analisado com LSA (do inglês, *Latent Semantic Analysis*). Na integração um modelo linear com pesos para cada característica foi aplicado. Como esperado, a técnica multimodal teve melhor desempenho em todos os noticiários e técnicas de integração com SVM foram comparadas com a abordagem desenvolvida, obtendo desempenhos muito semelhantes. As características descritas, assim como os resultados descritos, tornam esse trabalho o mais completo até o momento.

Um problema muito comum em todos estes trabalhos que representam o estado da arte de segmentação de cenas em telejornais é a falta de uma definição mais geral para cenas, uma vez que existem três tipos (notícias, vinhetas e comerciais) e várias maneiras diferentes de transição entre elas. Um exemplo é quando um âncora relata a notícia sem aparecer na imagem da TV, apenas narrando os acontecimentos. Desse modo, algumas perguntas ficam em aberto: *No início, quando são apresentadas resumidamente as notícias, elas são consideradas na avaliação dos resultados?*; *As vinhetas fazem parte de alguma transição de notícias?*; *Quando não há imagem do âncora mas uma notícia é apresentada, esta é considerada?*; *As transições entre os blocos de notícias e os comerciais são consideradas?*. Portanto, fica evidente que é necessária uma apresentação conceitual mais abrangente de cenas e suas transições, pois sem isso fica difícil analisar o desempenho das técnicas desenvolvidas pelos autores. Outro ponto não relacionado nos trabalhos multimodais é o uso dos símbolos do *closed-captions* para auxiliar na identificação de transição de notícias, sendo que esse conteúdo indicam as falas dos âncoras e momento exato que isso ocorre.

Observa-se que nos trabalhos multimodais, *ranking* é abordado com frequência como técnica de integração das mídias, obtendo resultados expressivos quando empregado. Por fim, mesmo que amplamente utilizadas, Chua et al. (2004) citam que o uso de algoritmos de aprendizado de máquina nas técnicas multimodais não descobrem muitas cenas, por conta do não treinamento adequado dos dados.

3.5 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 et al. (2003) descrevem 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 3.4 apresenta a estrutura hierárquica utilizando a semântica do MPEG-7.

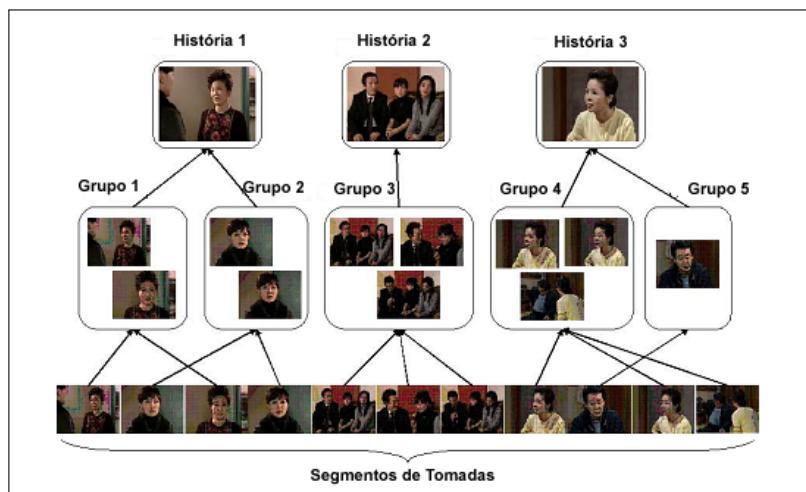


Figura 3.4: Estrutura hierárquica de representação de cena com MPEG-7 (Lee et al., 2003)

O uso de ontologias é outra abordagem recente para representar dados multimídia de uma maneira mais organizada, visando recuperação semântica facilitada. De acordo com a comunidade de Inteligência Artificial, “ontologia é uma especificação formal de conceitualização” (Gruber, 1993). A conceitualização envolvida no contexto dessa pesquisa refere-se ao domínio de conhecimento associado às características de imagens. Assim, 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. (2007b) relatam 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 semânticas pré-definidas (Figura 3.5) auxilia o usuário no sentido de permitir que este possa selecionar facilmente palavras-chaves para formular uma busca (Fan et al., 2008b,a).

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).

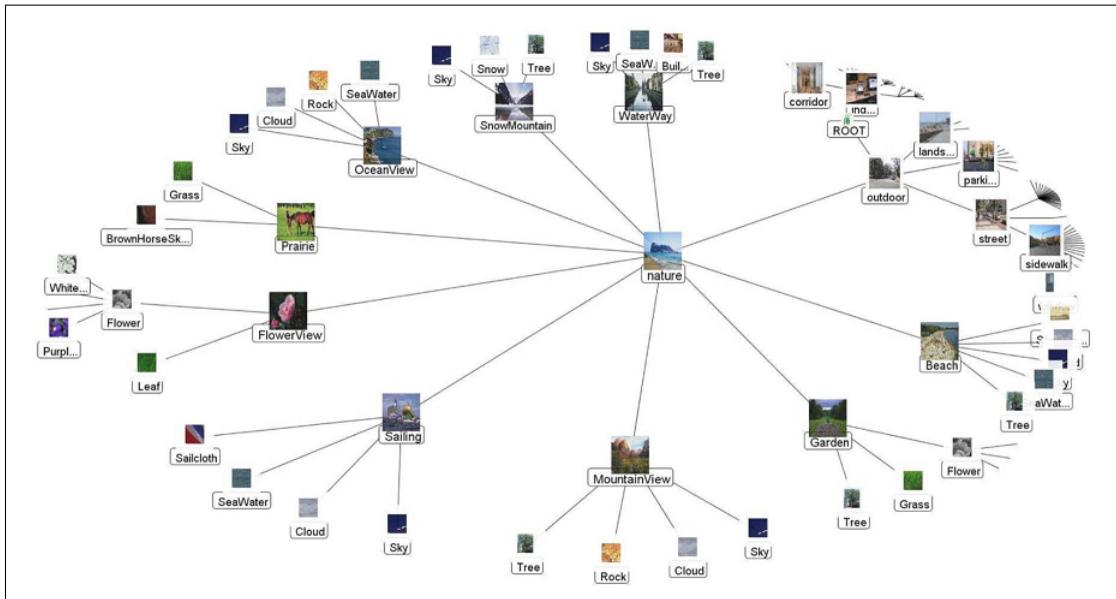


Figura 3.5: Visualização da ontologia com categorias pré-definidas (Fan et al., 2008b)

Tecnologias de compressão mais sofisticadas, como o MPEG-4 (for Standardisation, 2002), também possui estudos na área de segmentação semântica. Cavallaro et al. (2003) desenvolveram 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 recuperação de cenas 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., 2007b; Deb & Zhang, 2004).

A extração de texto de vídeos é uma metodologia que pode ajudar na classificação de vídeos. (Manzato & Goularte, 2008) realizam a comparação de técnicas como algoritmos genéticos e índice semântico de latência (do inglês, *Latent Semantic Indexing-(LSI)*) 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 polissemia⁴, que origina falsos positivos na classificação e o pequeno volume de texto empregado na amostra da metodologia. *Closed-captions* continuou a ser utilizado por Manzato et al. (2010) para identificar cenas em telejornais. Nesse trabalho os autores desenvolveram uma técnica que considera o usuário como produtor e fornecedor de conteúdo. Por meio de um mecanismo de busca, um arcabouço considera a combinação de diferentes critérios de pesquisa, como: características visuais (histograma local de cor) baseada em amostras de imagens (quadros-chave), texto obtido por intermédio de *closed caption* e reconhecimento de faces. Um algoritmo que faz a união desses critérios foi proposto, obtendo melhores resultados do que o uso das técnicas em separado, melhorando a experiência final do usuário com esse tipo de conteúdo.

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 et al., 2007a). Na categoria supervisionado, SVM, classificador Bayesiano (Jin et al., 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 et al., 2004) e Corte Normalizado (do inglês, *Normalized Cut- NCut*) (Ng et al., 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 (Vasconcelos, 2004) para classificação por probabilidade.

3.6 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, sonoras e textuais ou combinação delas, como pode ser observado também no apêndice deste trabalho, é uma tendência que visa conseguir melhores resultados.

Outro ponto importante nos trabalhos é a especificação de um determinado gênero do vídeo para aplicar as suas respectivas metodologias, tornando mais fácil a extração das cenas. Em alguns gêneros, como os telejornais, as pesquisas avançam consideravelmente em direção à multimodalidade das técnicas, entretanto, apresentam problemas em utilizar uma definição mais geral para os segmentos analisados (cenas). A abordagem de identificação do âncora é o principal indício do início de uma cena, mas isso não é regra, visto que dois âncoras ou mesmo uma imagem sem a figura de quaisquer âncora(s) pode indicar

⁴Polissemia são palavras que possuem mais de um significado.

o início de uma cena. Logo, histogramas de cor são usados para identificar a imagem de um âncora com determinado plano de fundo, seja ele estático (com uso de histogramas globais) ou dinâmico (com uso de histogramas locais), mas falha em detectar situações em que ocorrem mais de um âncora na mesma imagem. Para contornar este problema, reconhecimento de faces é uma abordagem que consegue bons resultados, no entanto, não resolve a questão de uma cena iniciar sem a figura de âncora algum, no caso em que somente a fala do âncora indica a transição de cena. Uma alternativa a essa situação é recorrer às características de áudio e texto. Portanto, faz-se necessário desenvolver um conjunto de técnicas multimodais que detectem todas as possíveis situações de transições de cenas.

Com os trabalhos apresentados neste capítulo, observa-se que as 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 ocasionado pelos trabalhos que realizam a recuperação de informação em vídeos digitais.

Segmentação Multimodal de Cenas: uma proposta

4.1 Considerações Iniciais

Este capítulo apresenta técnicas que extraem informações provenientes do conteúdo do vídeo com o propósito de segmentar estruturas denominadas cenas por meio da identificação do momento em que ocorrem suas transições.

As técnicas empregadas neste trabalho serão apresentadas em detalhes e de acordo com a mídia utilizada. Posteriormente, será apresentada a técnica multimodal proposta, assim como os resultados das técnicas e aplicadas em separado, com os resultados da técnica multimodal. Por fim, os resultados são comparados e analisados, destacando vantagens e desvantagens de cada abordagem.

4.2 Técnicas Empregadas

Foram empregadas ao todo sete técnicas com o intuito de extrair informações das mídias que compõem o vídeo, três delas atuando na recuperação de informação em imagens, duas analisando o fluxo de áudio, uma na parte textual e desenvolvida uma última integrando todas.

A seguir serão descritas as metodologias associadas às técnicas desenvolvidas, as quais analisam, cada uma, um tipo particular de mídia. Para avaliar os resultados obtidos, todas as técnicas produzem como saída constantes associadas à linha temporal do vídeo, especificamente carimbos de tempo (do inglês, *timestamps*). Os *timestamps* são um modo de representar essa linha temporal, tornando-se indispensáveis na integração de multimo-

dalidades, ou seja, na sincronização e alinhamento dos diferentes tipos de mídias (Eickeler & Muller, 1999; Huang et al., 1999; Alatan et al., 2001; Snoek & Worring, 2005).

4.2.1 Extração de Informação em Imagens

As metodologias abordadas nesta seção visam extrair informações nas imagens obtidas pelos vídeos. Todas as técnicas de imagem atuam e consideram em sua implementação o espaço de cor RGB (do inglês, *Red-Green-Blue* - Vermelho-Verde-Azul). Para realizar a identificação de transição de cenas, as seguintes etapas foram seguidas, por ordem:

1. Extração de quadros a cada 1 segundo do fluxo de vídeo.
2. Construção de uma base de dados de imagens para cada vídeo, identificando todas as tomadas por meio da extração do primeiro quadro de cada tomada como seu quadro mais representativo (Li et al., 2001).
3. Aplicação de uma técnica de extração de característica visual na base de imagens. Três técnicas foram utilizadas neste trabalho: histograma de cor (Fórmula 2.3) global e local e wavelets (Fórmula 2.9).
4. Aplicação de uma medida de similaridade na base de imagens, utilizando a primeira imagem do âncora como modelo de busca (por exemplo, a Figura 4.1). Para os histogramas utilizou-se a medida de similaridade denominada intersecção de histogramas (Fórmula 2.4 para histograma global e Fórmula 2.5 para histograma local) e para a wavelet utilizou-se a distância euclidiana (Fórmula 2.10). Enquanto a intersecção de histogramas retorna um valor no intervalo de 0 e 1 (quanto mais próximo de 1, mais semelhante é a imagem), a distância euclidiana calcula a diferença dos valores das assinaturas retornadas pela wavelet de cada imagem, considerando que quanto menor o valor, mais semelhante é a imagem.



Figura 4.1: Imagem modelo do âncora para busca

5. Extração dos *timestamps* dos m quadros/imagens mais semelhantes, por meio do tempo (em segundos) que a imagem foi extraída do vídeo. Isso é possível porque na etapa de extração dos quadros dos vídeos a cada segundo, a imagem gerada foi nomeada usando o tempo da captura. Por exemplo, se no nome da imagem consta

“000045”, este nome implica que o *timestamp* da imagem, considerando o formato hh:mm:ss, é 00:00:45 (ou seja, zero hora, zero minuto e quarenta e cinco segundos).

4.2.2 Extração de Informação em Áudio

Foram desenvolvidas duas metodologias de extração de informação de áudio para detectar transições de cenas. Relativo à captura das faixas de áudio correspondente aos vídeos, essa etapa foi efetuada com o auxílio da ferramenta de código aberto VirtualDub¹. Ambas metodologias utilizam técnicas que visam a detecção de momentos de silêncio na faixa de áudio e são descritas a seguir.

4.2.2.1 Audacity

Para efetuar a captura dos momentos de silêncio foi utilizada, para a análise e manipulação das faixas de áudio, a ferramenta Audacity². A faixa de áudio do vídeo é inserida no programa, o qual tem uma funcionalidade denominada *Silence Finder*, que encontra momentos de silêncio no vídeo com base em dois parâmetros: tempo (medido em segundos) e nível de ruído do sinal (medido em decibéis - dB). Dessa maneira, o algoritmo da ferramenta pode ser representado usando a função A , para qualquer clipe c_i , dada por:

$$A(c_i) = \begin{cases} \text{transição de cena: } r_i \leq R \\ \text{sem transição: caso contrário} \end{cases} \quad (4.1)$$

onde R é um limiar associado ao nível de ruído do áudio (medido em dB) e r_i é o nível de ruído associado à c_i , com $i = 1, \dots, n \in \mathbb{N}$. Os valores, definidos empiricamente, foram de 15 dB para R e de 1,1 segundo para cada clipe c , ou seja, o algoritmo da ferramenta detecta silêncio para clipes de, no mínimo, 1,1 segundo com 15 dB ou menos.

A Figura 4.2 apresenta parte da interface visual do Audacity para a representação da faixa de áudio, juntamente com os indicadores de silêncio, representados por linhas azuis na vertical acompanhadas de quadrados com a letra S em seus interiores.

Por se tratar de uma ferramenta de código aberto, a tarefa de automatizar a extração dos resultados, ou seja, os períodos de silêncio, foi realizada. Por conseguinte, alterando o código da ferramenta, executou-se de modo automático a extração dos *timestamps* que correspondem aos momentos de silêncio das faixas de áudio contidos nos telejornais analisados. Logo, para cada telejornal criou-se um arquivo de texto contendo os *timestamps* obtidos como saída de dados da ferramenta.

¹<http://www.virtualdub.org>

²<http://audacity.sourceforge.net/>

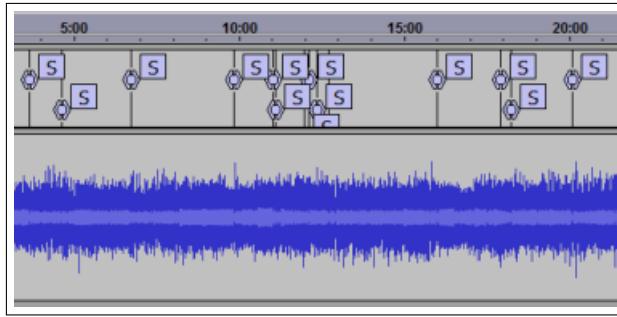


Figura 4.2: Parte da Interface visual da ferramenta Audacity

4.2.2.2 *Root Mean Square* - RMS

Também com o objetivo de detectar silêncio nas faixas de áudio dos vídeos, foi desenvolvido um sistema em Java, com o auxílio da biblioteca nativa *javax.sound* que calculasse a raiz da média dos quadrados (do inglês, *Root Mean Square*-RMS), revelando a variação temporal da magnitude de um sinal em relação à distribuição do volume nos clipes de áudio. O cálculo é feito, como o próprio nome diz, por um conjunto de quadros do som e computando a raiz quadrada da soma dos quadrados dos valores das amostras destes quadros, representado por:

$$v(n) = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} s_n^2(i)} \quad (4.2)$$

onde N denota o tamanho do quadro, s_i denota a i -ésima amostra no n -ésimo quadro, com i e $n \in \mathbb{N}$. Neste estudo, dividiu-se o fluxo de áudio em 1 quadro por segundo e para cada quadro obteve-se 48000 amostras. O cálculo do RMS é aplicado em cada quadro, ou seja, nas 48.000 amostras e, caso seja menor que um limiar (definido empiricamente como 0.5), é detectado silêncio. Assim, para cada quadro detectado com silêncio, extraiu-se o *timestamp* do mesmo baseado na sua posição, ou seja, se no quadragésimo quinto quadro detectou-se silêncio, então seu *timestamp* é 00:00:45. Na última etapa, criou-se um arquivo de texto para cada telejornal contendo os *timestamps* obtidos como saída de dados da técnica.

É importante observar que o volume do sinal do áudio depende da maneira como é efetuada a gravação e a digitalização do áudio do sistema, ou seja, é importante que, as faixas de áudio utilizadas no sistema sejam gravadas em um mesmo nível de volume e não sofram quaisquer alterações em suas análises (Liu et al., 1998b).

4.2.3 Extração de Informação em Texto

O modo de extração de informação de texto adotado foi utilizando o conteúdo menos explorado por trabalhos da área, os *closed-captions* dos vídeos. Assim como reportado por Boggs & Petrie (2000), segmentos no texto indicam transições de determinados momentos de fala, particularmente os caracteres “>>”. Tais caracteres, nos telejornais brasileiros, indicam o momento em que o repórter apresenta uma notícia, seja ele o âncora ou o próprio repórter produtor da notícia.

Após o processo de captura, foi necessário utilizar ferramentas e aplicar algoritmos para realizar a extração das informações necessárias no arquivo. Inicialmente, foi gerado um conjunto de arquivos SubRip (.srt) usando a ferramenta de código aberto CCExtractor³, a qual convertia o *closed-caption* binário, capturado com o Microsoft Graph Edit⁴ das fontes analógicas, para texto em formato de legenda.

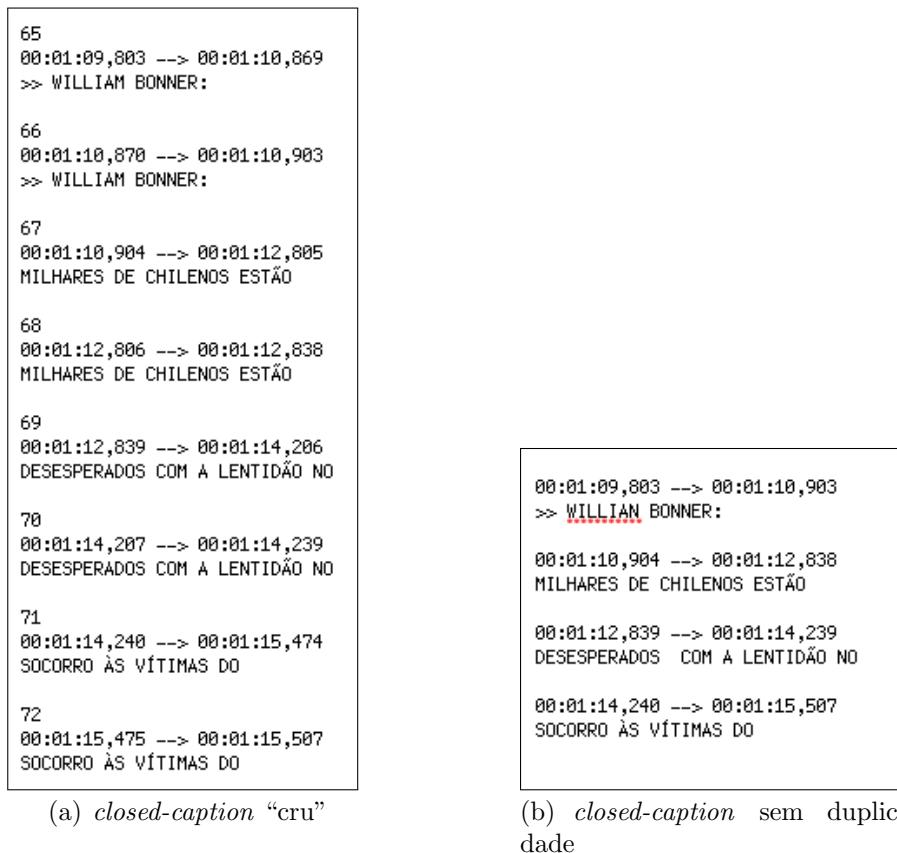


Figura 4.3: Representações do *closed-caption*

Um exemplo de *closed-caption* em sua forma “crua”, ou seja, assim que é capturado, é apresentado na Figura 4.3(a).

Por fim, foram desenvolvidos, em linguagem Java, e aplicados dois algoritmos a fim de obter as informações de transições de cenas por meio das falas dos âncoras do telejor-

³<http://ccextractor.sourceforge.net>.

⁴[http://msdn.microsoft.com/en-us/library/dd390950\(VS.85\).aspx](http://msdn.microsoft.com/en-us/library/dd390950(VS.85).aspx).

nal. O primeiro algoritmo visa retirar duplicidades existentes neste formato, ocasionando textos sem repetição de falas, conforme o texto ilustrado na Figura 4.3(b), obtido após aplicar o algoritmo 4.1 no texto da Figura 4.3(a). Resumidamente, este algoritmo verifica se os textos consecutivos são iguais e, caso sejam, utiliza-se o tempo início do primeiro texto, o tempo final do segundo e remove-se o segundo texto, caso contrário não realiza nenhuma operação e continua a verificação dos textos. O detalhe deste algoritmo é o método `TempoSubstringInicio(0,12)` e `TempoSubstringFim(17,29)`, que correspondem, respectivamente, à captura da string do tempo inicial e final do texto, as quais têm o seguinte formato: `00:00:00,000 --> 11:11:11,111`.

```

Entrada: Arquivo Texto T com closed-captions
Saída: Arquivo Texto TC sem os closed-captions redundantes
String linha, proxima_linha;
proxima_linha ← T.le_linha();
linha ← proxima_linha;

enquanto linha diferente de nulo faz
    linha = T.le_linha();
    proxima_linha ← T.le_linha();

    se linha.Texto igual a proxima_linha.Texto então
        TC.escreve(linha.TempoSubstringInicio(0,12) + " --> " +
        proxima_linha.TempoSubstringFim(17,29));
        nova_linha();
        TC.escreve(linha.Texto);
        nova_linha();
    fim
    senão
        repita
            TC.escreve(linha.TempoSubstringInicio(0,12) + " --> " +
            linha.TempoSubstringFim(17,29));
            TC.escreve(nova_linha());
            TC.escreve(linha.Texto);
            TC.escreve(nova_linha());

            linha.TempoSubstringInicio ← proxima_linha.TempoSubstringInicio;
            linha.TempoSubstringFim ← proxima_linha.TempoSubstringFim;
            linha.Texto ← proxima_linha.Texto;
            proxima_linha ← T.le_linha();

        até linha.Texto diferente de proxima_linha.Texto;
    fim
    TC.escreve(linha.TempoSubstringInicio(0,12) + " --> " +
    proxima_linha.TempoSubstringFim(17,29));
    TC.escreve(nova_linha());
    TC.escreve(linha.Texto);
    TC.escreve(nova_linha());
fim
retorna TC;

```

Algoritmo 4.1: Algoritmo para retirar as redundâncias do *closed-caption*

O segundo e último (Algoritmo 4.2) visa identificar o momento em que ocorrem as falas dos âncoras, as quais representam os inícios de cada cenas de notícias. Tanto as falas do(s) âncora(s) quanto a dos repórteres produtores das notícias são representadas no texto pelo símbolo “>>”, todavia, quando as falas são dos âncoras, o símbolo é sucedido com o nome do âncora (*e.g.*: “>> Fatima Bernardes: Boa Noite”), já quando a fala são dos repórteres não aparece o nome do mesmo, somente a palavra “repórter” (*e.g.*: “>> repórter: O sistema financeiro...”). Portanto, este algoritmo faz a análise do texto, procurando por linhas que possuam o símbolo mencionado e que não tenham a palavra repórter nem o carácter “:”, pois estes dois termos caracterizam a fala dos repórteres.

```

Entrada: Arquivo Texto TC com closed-captions sem redundâncias
Saída: Arquivo Texto TS somente com os timestamps da fala dos âncoras

String linha, proxima_linha;
proxima_linha ← T.le_linha() ;
linha ← proxima_linha ;

enquanto linha diferente de nulo faça
    linha ← T.le_linha();
    se linha.começaCom “>>” e (linha.nãoContem “repórter” e linha.nãoContem “:” )
    então
        proxima_linha ← proxima_linha.TempoSubstringInicio(0,12);
        TS.escreve(proxima_linha);
    fim
    proxima_linha ← T.le_linha() ;
    se proxima_linha.começaCom “>>” e (proxima_linha.nãoContem “repórter” e
    proxima_linha.nãoContem “:” ) então
        linha ← linha.TempoSubstringInicio(0,12);
        TS.escreve(linha);
    fim
fim
retorna TS;
```

Algoritmo 4.2: Algoritmo para capturar o *timestamp* das falas do(s) âncoras

4.3 Técnica Multimodal Proposta

O desenvolvimento de técnicas que atuam em mais de uma mídia na identificação de estruturas temporais de vídeo, também chamadas de técnicas multimodais (Snoek et al., 2005), são cada vez mais frequentes nos trabalhos científicos da área de recuperação de conteúdo multimídia (Ngo et al., 2001; Hanjalic, 2004; Li et al., 2004; Coimbra & Goularte, 2009; Manzato et al., 2009, 2010). Como mencionado no Capítulo 3, especificamente nas Subseções 3.4 e 3.5, agrupar e integrar técnicas, sejam elas relacionadas às características de imagem e/ou áudio e/ou texto, aumentam a quantidade de informações

sobre o conteúdo, proporcionando melhorias significativas nos resultados da segmentação, principalmente quando esses segmentos possuem uma carga maior de semântica, como acontecem com as cenas.

Consequentemente, a proposta deste trabalho foi desenvolver uma técnica multimodal que contenha informações de todas as mídias presentes no vídeo, conforme observado na Figura 4.4. Os tipos de informações e as técnicas foram detalhadas neste capítulo, mas basicamente, o fluxo de vídeo é dividido em três grupos, cada qual contendo informações de uma determinada mídia. No grupo relacionados às características visuais as informações analisadas foram obtidas das imagens (quadros-chaves), no grupo de características sonoras as informações foram capturadas do fluxo de áudio e a informação textual foi obtida do *closed-caption*. Após, aplicou-se a técnica multimodal integrando os resultados das técnicas e, por fim, as transições de cenas são obtidas, lembrando que tanto os resultados das técnicas em separado quanto o resultado da integração delas (técnica multimodal) são representadas por *timestamps*.

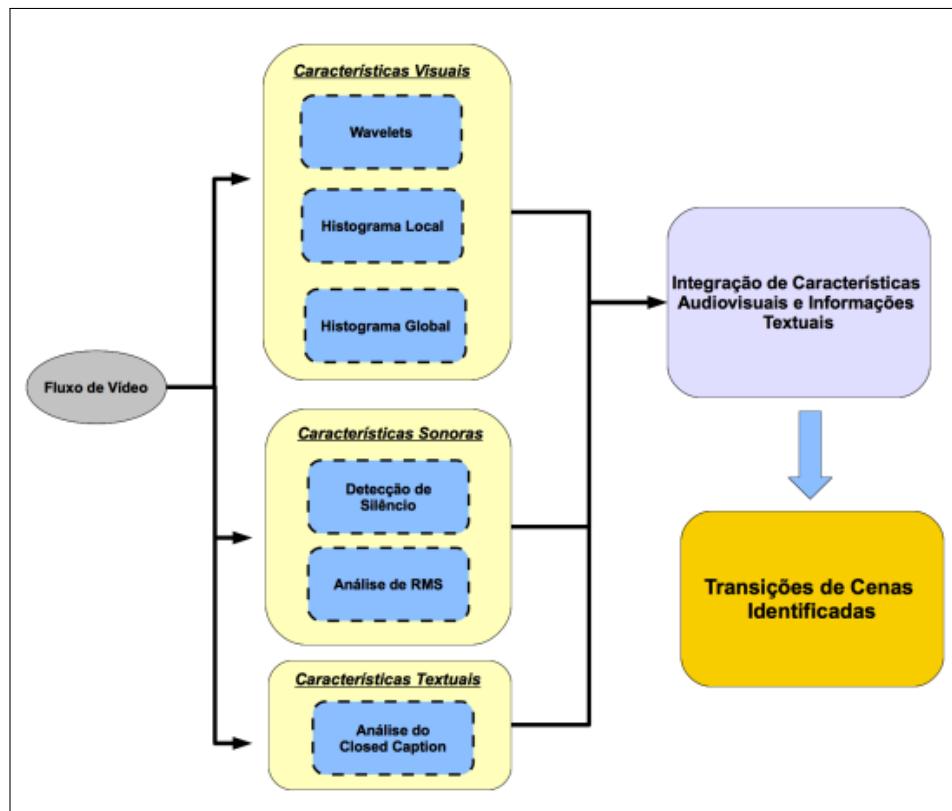


Figura 4.4: Arquitetura da técnica multimodal

De modo geral, a técnica multimodal faz uso de uma tabela de espalhamento (do inglês, *hashing table*) que possui seus valores baseados em um *ranking* associado à pesos pré-definidos para cada uma das técnicas. Deste modo, é possível verificar se uma determinada transição de cena foi identificada por outras técnicas garantindo mais confiança e eficiência ao resultado.

A tabela de espalhamento em questão é composta por chaves de pesquisas e valores de acordo com a seguinte definição:

Definição 1. Considera-se $C = c_1, c_2, \dots, c_n$ um conjunto de chaves, onde c_i é um identificador para uma transição de cena, no caso adotou-se o timestamp, com $i = 1, \dots, n \in \mathbb{N}$. Como não há 2 cenas em um mesmo instante de tempo, a chave é única. Considera-se também $V = v_1, v_2, \dots, v_n$ um conjunto de valores, onde $v_i = rank_{comb} \in \mathbb{R}^+ \leq 1$ (é um número real positivo menor ou igual a 1), que é calculado de acordo com o ranking de agregação de resultados. Define-se S como um conjunto de pares chaves valor (c_i, v_i) onde $c_i \in C$ e $v_i \in V$.

O ranking das cenas de um determinado vídeo são formados pela soma de pesos, os quais variam de técnica para técnica. A abordagem para a definição dos valores dos pesos segue a eficácia de cada técnica, a ser apresentada na Tabela 4.1 da próxima Subseção (4.4), ou seja, quanto maior a precisão e revocação da técnica, maior seu peso. Assim, a proposta de integração neste trabalho é definida como:

$$rank_{comb}(v_i) = \begin{cases} \text{se } T_q \in T_{cc_texto} \\ \quad valor(v_i) = valor(v_i) + 0.3 \\ \\ \text{se } T_q \in T_{audacity_audio} \\ \quad valor(v_i) = valor(v_i) + 0.1 \\ \\ \text{se } T_q \in T_{rms_audio} \\ \quad valor(v_i) = valor(v_i) + 0.1 \\ \\ \text{se } T_q \in T_{hglobal_img} \\ \quad valor(v_i) = valor(v_i) + 0.2 \\ \\ \text{se } T_q \in T_{hlocal_img} \\ \quad valor(v_i) = valor(v_i) + 0.1 \\ \\ \text{se } T_q \in T_{wavelet_img} \\ \quad valor(v_i) = valor(v_i) + 0.2 \end{cases} \quad (4.3)$$

onde $rank_{comb}(v_i) \in V$, com $i = 1, \dots, n$; T_q é o timestamp a ser analisado.

$T_{cc_texto|audacity_audio|rms_audio|hglobal_img|hlocal_img|wavelet_img}$ corresponde, respectivamente, ao timestamp do closed-caption(texto), Audacity (áudio), RMS (áudio), Histograma Local (imagem), Histograma Global (imagem) e Wavelets (imagem).

Considere que T seja um arquivo texto que contenha os timestamps obtidos manualmente do vídeo. Basicamente, uma tabela de espalhamento R é formada por chaves que

correspondem somente aos *timestamps* detectados por alguma das técnicas quando comparados com os *timestamps* de T . Desta maneira, para cada *timestamp* de uma técnica, é verificado se o mesmo existe em T e, em caso afirmativo, é adicionado à chave de R o *timestamp*, e também o seu respectivo valor, o qual é calculado somando o peso da técnica ao seu antigo valor.

Toda abordagem que realiza a extração de *timestamp* com o intuito de integrar mais de uma técnica, deve tratar da questão do alinhamento destes tipos de dados, principalmente quando é necessário uma comparação entre eles. Isso acontece porque no processo de captura, independente da mídia, pode ocorrer arredondamento dos milissegundos pra mais ou pra menos, ocasionando uma diferença de segundos. Por exemplo, duas técnicas distintas detectam corretamente os seguintes *timestamps*: “00:35:93” e “00:35:91”, sabendo que há uma transição de cena em “00:35:92”. Neste caso, o processo de captura da primeira técnica arredondou um segundo para cima e a segunda um segundo para baixo. Portanto, foi desenvolvido um algoritmo que considerasse um intervalo de tempo quando comparado os tempos do arquivo modelo com o do obtido por alguma técnica (Algoritmo 4.3).

Entrada: Arquivo Texto A com os *timestamps* corretos (obtidos manualmente)

Entrada: *Timestamp* T de alguma técnica a ser analisada

Saída: Valor booleano que indica se os tempos pertencem à mesma transição

```

String tempo, tempoMin, tempoMax;
Inteiro valor;
Boleano transicao;
linha ← A.le_linha();
transicao ← falso;
enquanto linha diferente de nulo faça
    tempo ← A.le_linha() ;
    tempoMax ← tempo.adicionaSegundos( +1,5);
    tempoMin ← tempo.adicionaSegundos( -1,5) ;
    se  $T$  antes de tempoMax e  $T$  depois de tempoMin então
        |   transicao ← verdadeiro;
    fim
fim
retorna transicao;
```

Algoritmo 4.3: Algoritmo usado na comparação dos tempos

O algoritmo 4.3 verifica se no arquivo com os tempos corretos criado manualmente há algum *timestamp* que indique a mesma transição de cena de um determinado *timestamp* (T) de uma técnica. Na comparação entre os tempos, é verificado se T está no intervalo de um segundo e meio a mais ou a menos para todos os tempos do arquivo A. Em caso afirmativo, verifica-se a qual técnica o *timestamp* T pertence a fim de calcular o valor dos pesos, adicionando-o na tabela de espalhamento.

Posteriormente à etapa de *ranking* e ao algoritmo de sincronização dos tempos, são selecionadas as m transições de cenas melhores rankiadas, sendo $m \leq n$.

4.4 Resultados Obtidos

Esta seção descreve os resultados obtidos das técnicas aplicadas em uma base de vídeos específica e está assim dividida:

- Ambiente de Testes
- Técnica Textual
- Técnicas Sonoras
- Técnicas Visuais
- Técnica Multimodal

Mesmo que apresentados e discutidos adiante e em detalhes, a Tabela 4.1 sumariza os resultados de todas as técnicas para melhor comparação entre eles. Nesta tabela os resultados são apresentados conforme as medidas de avaliação precisão e revocação, representadas respectivamente por P e R e apresentadas no formato de porcentagem (%). Cc é a abreviação para *closed-caption*, o qual representa a técnica de extração de informação de texto. Audacity e RMS correspondem, respectivamente, a ferramenta utilizada e a técnica *Root Mean Square* de extração de informação de áudio, ambas detectando silêncio. Os valores 30 e 15 representam o valor de m para as m transições de cenas melhores rankiadas, seja para as técnicas de extração de informação das imagens ou para a técnica multimodal. Por fim, \bar{x} indica a média aritmética simples dos resultados de cada técnica para cada medida de avaliação, representada por:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (4.4)$$

em que x_i representa o resultado de precisão ou revocação de uma técnica em uma edição de um telejornal e n indica a quantidade de telejornais analisados, ou seja, 20.

Para calcular o desempenho (precisão e a revocação) de cada técnica, são comparados dois conjuntos de *timestamps*. Considere C1 o primeiro conjunto de *timestamps* obtido como saída de cada técnica e o segundo conjunto C2 relativo às transições reais de cenas, efetuado manualmente. Logo, aplica-se um algoritmo comparando os *timestamps* de C1 e C2 semelhante ao Algoritmo 4.3, tendo como única diferença o intervalo de tempo, o qual é de 1 segundo (1 segundo para mais ou para menos). Desse modo, quanto mais *timestamps* similares (considerando o intervalo de 1 segundo) de C1 em C2, melhor é o desempenho da técnica, ou seja, maiores os valores de precisão e revocação.

Tabela 4.1: Resultados das técnicas

Texto	C_c	Áudio				Imagen																
		Hist. Global				Hist. Local				Wavelet												
		30		15		30		15		30		15		30		15						
P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R					
JN 22-02	75.0	57.1	54.2	31.0	28.0	33.3	43.3	31.0	28.9	43.3	31.0	60.0	26.7	50.0	35.7	66.7	60.0	55.6	66.7	31.1		
JN 02-03	76.5	60.5	44.1	34.9	33.3	44.2	70.0	48.9	73.3	26.2	53.3	37.2	86.7	21.4	83.3	58.1	80.0	23.1	86.7	42.9	86.7	23.8
JN 03-03	77.4	53.3	59.5	48.9	25.6	48.9	55.3	35.6	80.0	25.6	46.7	31.1	73.3	30.2	70.0	46.7	80.0	27.9	80.0	60.5	100	30.2
JN 04-03	86.1	63.2	50.0	40.9	41.3	53.1	66.7	40.8	66.7	26.7	53.3	32.7	66.7	24.4	60.0	36.7	60.0	26.7	86.7	53.3	83.3	33.3
JN 09-03	69.0	44.4	64.5	44.4	40.0	48.9	63.3	42.2	86.7	21.3	46.7	31.1	80.0	21.3	50.0	33.3	80.0	19.1	83.3	55.3	93.3	27.7
JH 02-03	72.7	22.3	27.2	8.6	26.7	22.9	63.3	54.3	80.0	34.3	50.0	42.9	80.0	34.3	60.0	51.4	73.0	31.4	70.0	60.0	86.7	37.1
JH 03-03	64.9	53.3	36.4	8.9	25.0	15.6	43.3	28.9	60.0	20.0	40.0	26.7	73.3	24.4	53.3	35.6	73.0	24.4	60.0	40.0	86.7	28.9
JH 05-03	69.2	22.0	28.6	4.9	27.3	14.7	46.7	34.1	93.3	34.1	46.7	34.1	93.3	34.1	46.7	34.1	93.3	34.1	70.0	51.2	80.0	29.3
JH 06-03	87.5	19.4	47.4	25.0	29.8	38.9	63.3	52.8	93.3	38.9	63.3	52.8	93.3	36.1	70.0	58.3	86.7	36.1	76.7	63.9	86.7	36.1
JH 10-03	54.2	36.1	36.8	19.4	28.1	25.0	73.3	61.1	86.6	36.1	67.7	58.3	80.0	33.3	73.3	61.1	73.0	30.6	73.3	61.1	80.0	33.3
JR 16-07	52.7	69.6	37.2	28.6	18.5	30.4	46.7	25.0	73.3	19.3	56.7	30.4	66.7	17.5	33.3	17.9	66.7	17.5	70.0	36.8	86.7	22.8
JR 19-07	50.0	73.1	42.2	36.5	16.0	37.0	53.3	30.8	80.0	23.1	43.3	25.0	73.3	21.2	46.7	26.9	66.7	19.2	70.0	40.4	80.0	23.1
JR 20-07	45.3	60.7	32.3	17.9	28.8	33.9	50.0	26.8	60.0	16.1	70.0	37.5	93.3	25.0	33.3	17.9	66.7	17.9	70.0	37.5	73.3	19.6
JR 21-07	29.6	70.8	14.3	18.8	14.2	43.8	70.0	43.8	86.7	27.1	54.8	35.4	80.0	25.0	53.3	33.3	80.0	25.0	73.3	45.8	93.3	29.2
JR 23-07	56.9	67.3	45.7	29.1	18.5	36.4	46.7	25.5	66.7	18.2	66.7	36.4	73.3	20.0	46.6	25.5	53.0	14.5	76.7	41.8	73.3	20.0
JG 04-03	93.8	42.9	32.5	37.1	20.0	37.1	66.7	57.1	86.7	37.1	63.3	54.3	80.0	34.3	53.3	45.7	80.0	34.3	90.0	77.1	93.3	40.0
JG 08-09	93.3	46.6	26.5	30.0	13.6	33.3	56.7	66.7	33.3	46.7	60.0	46.7	60.0	30.0	56.7	56.7	53.0	22.9	70.0	73.3	36.7	
JG 09-09	74.1	55.6	46.7	38.9	25.5	36.1	53.3	44.4	86.7	36.1	50.0	41.7	86.7	36.1	56.7	47.2	86.7	36.1	70.0	58.3	93.3	38.9
JG 21-09	64.0	48.5	40.0	24.2	27.3	36.4	60.0	54.5	100	45.0	20.0	18.2	100	45.5	70.0	63.6	93.3	42.4	63.3	57.6	86.7	39.4
JG 23-09	71.0	66.7	34.3	36.4	18.8	36.4	53.3	48.5	93.3	42.4	50.0	45.5	80.0	36.4	56.7	51.5	93.3	42.4	80.0	72.7	93.3	42.4
\bar{x}	68.2	51.7	40.0	28.2	25.3	35.3	57.2	42.1	79.7	29.5	51.7	37.4	79.0	28.9	56.2	41.9	75.3	27.6	74.0	54.1	84.8	31.1

4.4.1 Ambiente de Testes

As conhecidas bases de dados do meio acadêmico, como a TRECVID⁵ e a MediaMill⁶, não foram escolhidas para avaliar as técnicas. A primeira porque a base não é aberta à comunidade científica, somente aos participantes ativos da organização e a segunda em razão da insuficiente quantidade e qualidade de vídeos de telejornais e metadados correspondente aos mesmos, ou seja, não possuem uma quantidade suficiente de telejornais na íntegra e também não possuem os *closed-captions* associados aos mesmos.

Portanto, para aplicar as técnicas mencionadas foi criada uma base de vídeos composta por vinte telejornais brasileiros, sendo três noticiários da Rede Globo de Televisão (Jornal Nacional, Jornal Hoje, Jornal da Globo) e um noticiário da Rede Record de Televisão (Jornal da Record), cada noticiário contendo cinco telejornais/edições. No processo de captura dos vídeos foi utilizada a placa de captura de vídeo analógico de modelo PixelView PlayTV Cinema Pro, juntamente com a ferramenta VentiTV 1.0, a qual acompanha o dispositivo de captura. Os vídeos foram capturados usando a codificação MPEG-2, com resolução de imagem de 720 x 480 pixels, com taxa de quadros a 30 quadros por segundo e com áudio a 48 kHz de frequência. Informações mais detalhadas da estrutura temporal dos telejornais contendo a data de captura, tempo total de duração, quantidade de cenas e quantidade de tomadas são apresentadas na Tabela 4.2.

Tabela 4.2: Informações da estrutura temporal dos telejornais

Telejornal	Tempo(hh:mm:ss)	Tomadas	Cenas
Jornal Nacional 22-02-2010	00:38:41	518	45
Jornal Nacional 02-03-2010	00:39:00	545	42
Jornal Nacional 03-03-2010	00:34:27	470	43
Jornal Nacional 04-03-2010	00:51:17	601	45
Jornal Nacional 09-03-2010	00:38:59	499	47
Jornal Hoje 02-03-2010	00:29:16	407	35
Jornal Hoje 03-03-2010	00:34:58	447	45
Jornal Hoje 05-03-2010	00:34:08	492	41
Jornal Hoje 06-03-2010	00:32:09	968	36
Jornal Hoje 10-03-2010	00:33:18	414	36
Jornal da Record 16-07-2010	01:14:11	947	57
Jornal da Record 19-07-2010	01:13:50	918	52
Jornal da Record 20-07-2010	01:14:05	962	56
Jornal da Record 21-07-2010	01:12:46	804	48
Jornal da Record 23-07-2010	01:14:06	1.034	55
Jornal da Globo 04-03-2010	00:41:21	548	35
Jornal da Globo 08-09-2010	00:32:33	506	30
Jornal da Globo 09-09-2010	00:38:21	509	36
Jornal da Globo 21-09-2010	00:32:21	425	33
Jornal da Globo 23-09-2010	00:42:10	569	33

⁵<http://trecvid.nist.gov/>.

⁶<http://www.science.uva.nl/research/mediamill/>.

4.4.2 Técnica Textual

Os resultados obtidos com as técnicas de extração de informação textual, apresentados na Tabela 4.3, foram os melhores quando comparados somente com as outras técnicas aplicadas em separado. Isso acontece porque na formatação do texto contido no *closed-caption*, fica explícito a fala do âncora, a qual, de modo geral caracteriza o início de cenas de notícias. Entretanto, os resultados não foram melhores porque nem toda cena começa com a fala do âncora, por exemplo, um âncora pode retornar a comentar a mesma notícia ou ainda dois âncoras podem apresentar a mesma notícia.

Tabela 4.3: Resultados da técnica de texto

	<i>Closed-caption</i>	
	P	R
JN 22-02	75.0	57.1
JN 02-03	76.5	60.5
JN 03-03	77.4	53.3
JN 04-03	86.1	63.2
JN 09-03	69.0	44.4
JH 02-03	72.7	22.3
JH 03-03	64.9	53.3
JH 05-03	69.2	22.0
JH 06-03	87.5	19.4
JH 10-03	54.2	36.1
JR 16-07	52.7	69.6
JR 19-07	50.0	73.1
JR 20-07	45.3	60.7
JR 21-07	29.6	70.8
JR 23-07	56.9	67.3
JG 04-03	93.8	42.9
JG 08-09	93.3	46.6
JG 09-09	74.1	55.6
JG 21-09	64.0	48.5
JG 23-09	71.0	66.7
\bar{x}	68.2	51.7

Observou-se também que as informações textuais representaram com muita exatidão o conteúdo sonoro do telejornal, ou seja, quase que a totalidade das falas foram capturadas pelo *closed-caption*, inclusive respeitando um certo padrão, pois todos os telejornais apresentavam o símbolo “>>” para as falas do(s) âncora(s) e dos repórteres, com a palavra repórter acentuada. A exceção foi o telejornal Jornal Hoje, pois observou-se que em raros momentos do vídeo a fala não era capturada e em outros era omitido o símbolo que representava a fala do âncora.

A Figura 4.5 ilustra um trecho da edição do telejornal mencionado, notadamente datado de 06/03/2010, o qual contém os dois problemas citados. O primeiro pode ser

observado entre os instantes 00:01:40 e 00:01:49, os quais possuem falas de pessoas mas não há texto correspondente. O segundo problema ocorre na fala: “O jornal hoje deste sábado”, a qual deveria conter o seguinte (indicando a fala do âncora) “>>Evaristo Costa: O jornal hoje deste sábado”. O telejornal Jornal da Record também não apresentou a questão do símbolo indicando a fala do âncora em alguns poucos casos, ocasionando a queda de precisão em seus resultados.

```

00:01:28,483 --> 00:01:30,017
e amanhã também tem festa na

00:01:30,018 --> 00:01:33,020
laje.

00:01:40,328 --> 00:01:49,002
ô

00:01:54,008 --> 00:01:54,841
O jornal hoje deste sábado

00:01:54,842 --> 00:01:56,910
começa falando dos hábitos de

00:01:56,911 --> 00:01:57,577
consumo das mulheres.

```

Figura 4.5: Exemplo de um *closed-caption* mal formatado

Embora tenha os melhores resultados quando aplicada sozinha, com uma média aritmética de quase 70% para precisão e mais de 50% para revocação, a técnica de extração textual tem sua desvantagem e está associada à sua produção nos telejornais, pois nem todos os telejornais das principais emissoras possuem esse recurso e, quando possuem, nem sempre as informações textuais possuem 100% de qualidade fidedigna no que diz respeito a representação de todas as falas e de todos os símbolos, como aconteceu em algumas edições do telejornal Jornal Hoje e do Jornal da Record.

4.4.3 Técnicas Sonoras

Em relação às técnicas de áudio, que visam analisar períodos de silêncio nos telejornais, foi possível observar na Tabela 4.4 que os melhores resultados, tanto de precisão quanto de revocação, foram obtidos quando aplicaram-se ambas as técnicas na sequência de vídeos do telejornal Jornal Nacional. Baseado nessa observação fica evidente que as técnicas de RMS e a ferramenta Audacity possuem um desempenho melhor na sequência de vídeos do Jornal Nacional. Em contrapartida, o pior desempenho das técnicas está associado à sequência de vídeos do Jornal Hoje.

Contrapondo os resultados de áudio por meio da Tabela 4.4, percebe-se que a média da precisão da ferramenta Audacity foi melhor, com 40% de acertos, do que a técnica RMS,

que obteve uma média de cerca de 25%. Todavia, a revocação do RMS foi, em média, 7% melhor que a Audacity, ou seja, apesar do RMS ser menos preciso para detectar as transições de cenas, detectou-se uma quantidade maior do que a ferramenta Audacity.

Tabela 4.4: Resultados das técnicas sonoras

	Audacity		RMS	
	P	R	P	R
JN 22-02	54.2	31.0	28.0	33.3
JN 02-03	44.1	34.9	33.3	44.2
JN 03-03	59.5	48.9	25.6	48.9
JN 04-03	50.0	40.9	41.3	53.1
JN 09-03	64.5	44.4	40.0	48.9
JH 02-03	27.2	8.6	26.7	22.9
JH 03-03	36.4	8.9	25.0	15.6
JH 05-03	28.6	4.9	27.3	14.7
JH 06-03	47.4	25.0	29.8	38.9
JH 10-03	36.8	19.4	28.1	25.0
JR 16-07	37.2	28.6	18.5	30.4
JR 19-07	42.2	36.5	16.0	37.0
JR 20-07	32.3	17.9	28.8	33.9
JR 21-07	14.3	18.8	14.2	43.8
JR 23-07	45.7	29.1	18.5	36.4
JG 04-03	32.5	37.1	20.0	37.1
JG 08-09	26.5	30.0	13.6	33.3
JG 09-09	46.7	38.9	25.5	36.1
JG 21-09	40.0	24.2	27.3	36.4
JG 23-09	34.3	36.4	18.8	36.4
\bar{x}	40.0	28.2	25.3	35.3

De acordo com Zhang & Wang (2010), a detecção de silêncio, além de identificar transições de cenas em telejornais de um modo geral, considerando todos os tipos de notícias, este método também emergiu como um indicador confiável de presença de propagandas publicitárias, em outras palavras, este método também é utilizado para identificar transições de cenas em que ocorre somente a categoria de propagandas publicitárias. Portanto, baseada nessa afirmativa foi feita uma análise das técnicas com o intuito de avaliar seus desempenhos quando considerado somente as transições que ocorre propagandas publicitárias. A Tabela 4.5 apresenta a quantidade de transições que correspondem aos tipos de cenas mencionada de cada telejornal. É importante salientar que a coluna Início representa a transição que corresponde ao final do bloco do telejornal e ao início da propaganda, assim como o Fim representa o final da propaganda e início do bloco do telejornal. Logo, a coluna quantidade representa o número de transições, por exemplo, 10 transições correspondem a 5 transições de início e 5 de fim, o que é equivalente a dizer que o telejornal possui 5 intervalos de propagandas publicitárias.

Tabela 4.5: Resultados da identificação das transições de propagandas publicitárias com as técnicas de áudio

	Quantidade	Audacity		RMS	
		Início	Fim	Início	Fim
JN 22-02	8	3/4	4/4	2/4	4/4
JN 02-03	8	2/4	4/4	2/4	4/4
JN 03-03	6	1/3	3/3	2/3	3/3
JN 04-03	10	3/5	5/5	4/5	5/5
JN 09-03	8	3/4	4/4	2/4	4/4
JH 02-03	4	1/2	2/4	1/4	2/2
JH 03-03	4	0/2	2/2	0/2	2/2
JH 05-03	4	0/2	2/2	1/2	2/2
JH 06-03	4	0/2	2/2	1/2	2/2
JH 10-03	4	1/2	2/2	2/2	2/2
JR 16-07	4	2/2	2/2	2/2	2/2
JR 19-07	4	2/2	1/2	2/2	2/2
JR 20-07	4	2/2	1/2	2/2	2/2
JR 21-07	4	1/2	1/2	2/2	1/2
JR 23-07	4	2/2	1/2	2/2	2/2
JG 04-03	8	0/4	4/4	0/4	4/4
JG 08-09	6	0/3	2/3	0/4	3/3
JG 09-09	8	0/4	4/4	0/4	4/4
JG 21-09	6	0/3	3/3	1/4	3/3
JG 23-09	8	0/4	4/4	0/4	3/4
TOTAL		23/58	53/58	28/58	56/58
		76/116		84/116	

O principal resultado obtido na Tabela 4.5 está relacionado à identificação de transições que ocorrem no fim da propaganda e início do bloco do telejornal (coluna Fim). Tanto a ferramenta Audacity quanto a técnica RMS identificaram a grande maioria dessas transições, com destaque para RMS, identificando 56 das 58 existentes. Ao final da Tabela 4.5 é possível observar que RMS possui um desempenho melhor no geral (84/116) quando comparado com a ferramenta Audacity (76/116). Desse modo, embora com resultados regulares, as técnicas de áudio obtiveram resultados expressivos quando o intuito é identificar cenas do tipo propaganda publicitária (comercial), principalmente no momento em quem ela termina e é iniciado um novo bloco do telejornal.

4.4.4 Técnicas Visuais

Conforme os resultados apresentados na Tabela 4.6 envolvendo as características de imagem, o desempenho das técnicas de histograma global e wavelets foram similares para todos os grupos de telejornais, fato verificado também observando-se a média aritmética dessas técnicas, com uma pequena vantagem para o histograma global. Este resultado

contraria a tendência da literatura que apontam estudos, principalmente na área de CBIR, os quais as técnicas de wavelets fornecem melhores resultados que os histogramas para recuperação de imagens similares, entretanto, este fato acontece pelas peculiaridades da definição proposta para cena quando o tipo é notícia, a qual possui o âncora como modelo de imagem para efetuar buscas por similaridades.

Tabela 4.6: Resultados das técnicas visuais

	Hist. Global				Hist. Local				Wavelets			
	30		15		30		15		30		15	
	P	R	P	R	P	R	P	R	P	R	P	R
JN 22-02	43.3	31.0	73.3	28.9	43.3	31.0	60.0	26.7	50.0	35.7	66.7	26.7
JN 02-03	70.0	48.9	73.3	26.2	53.3	37.2	86.7	21.4	83.3	58.1	80.0	23.1
JN 03-03	53.3	35.6	80.0	25.6	46.7	31.1	73.3	30.2	70.0	46.7	80.0	27.9
JN 04-03	66.7	40.8	66.7	26.7	53.3	32.7	66.7	24.4	60.0	36.7	60.0	26.7
JN 09-03	63.3	42.2	86.7	21.3	46.7	31.1	80.0	21.3	50.0	33.3	80.0	19.1
JH 02-03	63.3	54.3	80.0	34.3	50.0	42.9	80.0	34.3	60.0	51.4	73.0	31.4
JH 03-03	43.3	28.9	60.0	20.0	40.0	26.7	73.3	24.4	53.3	35.6	73.0	24.4
JH 05-03	46.7	34.1	93.3	34.1	46.7	34.1	93.3	34.1	46.7	34.1	93.3	34.1
JH 06-03	63.3	52.8	93.3	38.9	63.3	52.8	93.3	36.1	70.0	58.3	86.7	36.1
JH 10-03	73.3	61.1	86.6	36.1	67.7	58.3	80.0	33.3	73.3	61.1	73.0	30.6
JR 16-07	46.7	25.0	73.3	19.3	56.7	30.4	66.7	17.5	33.3	17.9	66.7	17.5
JR 19-07	53.3	30.8	80.0	23.1	43.3	25.0	73.3	21.2	46.7	26.9	66.7	19.2
JR 20-07	50.0	26.8	60.0	16.1	70.0	37.5	93.3	25.0	33.3	17.9	66.7	17.9
JR 21-07	70.0	43.8	86.7	27.1	54.8	35.4	80.0	25.0	53.3	33.3	80.0	25.0
JR 23-07	46.7	25.5	66.7	18.2	66.7	36.4	73.3	20.0	46.6	25.5	53.0	14.5
JG 04-03	66.7	57.1	86.7	37.1	63.3	54.3	80.0	34.3	53.3	45.7	80.0	34.3
JG 08-09	56.7	56.7	66.7	33.3	46.7	46.7	60.0	30.0	56.7	56.7	53.0	22.9
JG 09-09	53.3	44.4	86.7	36.1	50.0	41.7	86.7	36.1	56.7	47.2	86.7	36.1
JG 21-09	60.0	54.5	100	45.0	20.0	18.2	100	45.5	70.0	63.6	93.3	42.4
JG 23-09	53.3	48.5	93.3	42.4	50.0	45.5	80.0	36.4	56.7	51.5	93.3	42.4
\bar{x}	57.2	42.1	79.7	29.5	51.7	37.4	79.0	28.9	56.2	41.9	75.3	27.6

As Figuras 4.7 e 4.8 ilustram, respectivamente, os 30 melhores resultados obtidos ordenados (de cima para baixo, da esquerda para a direita), das técnicas de histograma global e wavelets do telejornal Jornal Nacional de 09/03/2010, pois este representa um dos casos que o histograma global foi melhor que a wavelet. A técnica de histograma global, além de retornar mais imagens similares da imagem modelo (Figura 4.6), a qual é a imagem de um âncora, também retornou uma outra transição de cena verdadeira, onde tal imagem não aparece. Isso ocorre pela característica da técnica, pois o histograma global analisa a intensidade das cores da imagem como um todo (globalmente), de maneira que as imagens que possuem um plano de fundo semelhante acabam prevalecendo na similaridade, sendo identificadas. Deste modo, a imagem que representa a notícia de clima foi identificada porque há um plano de fundo semelhante entre as imagens dos âncoras e do plano de fundo da repórter da notícia da categoria clima. Em contrapartida, fica evidente nos resultados obtidos pela wavelet, para este mesmo telejornal, que a técnica

não identifica somente âncoras, mas também imagens com pessoas que estavam na mesma posição da imagem modelo do âncora, isso devido à característica da técnica que explora a textura da imagem, ressaltando o posicionamento espacial dos objetos da imagem.



Figura 4.6: Imagem modelo para busca no Jornal Nacional de 09/03/2010



Figura 4.7: Imagens resultantes da técnica de histograma global para o Jornal Nacional de 09/03/2010

Em relação ao histograma local, enquanto o global analisa a imagem como um todo, o local faz a análise de intensidade de cor por partes/blocos da imagem, possuindo assim uma característica intermediária entre o histograma global e wavelets, pois assim como as wavelets, considera a disposição espacial, não de objetos (textura), mas sim das cores. Com isso, o plano de fundo que no histograma global é mais importante, neste histograma fica com peso menor, ou seja, em segundo plano. Desse modo, os resultados do histograma local são os piores quando comparados com as outras duas técnicas, mas mesmo assim ainda são próximos. Cenas com ambos os âncoras juntos não foram retornadas por nenhuma técnica pelo fato deste telejornal não apresentar início de cena dessa



Figura 4.8: Imagens resultantes da técnica de wavelet para o Jornal Nacional de 09/03/2010

maneira.

O uso do histograma de cor mostrou-se satisfatório, porém foram poucos os casos que obteve melhores resultados quando comparado com outra técnica de extração de informação em image utilizado neste trabalho. Como apresentado na Figura 4.9, os resultados mostram que o algoritmo prioriza encontrar imagens de âncora da mesma pessoa da imagem modelo e somente encontrar todas elas, considera a imagem de outro pessoa como âncora. Isso ocorre porque o posicionamento espacial da cor é considerado na técnica. Outra vantagem obtida em outros telejornais foi considerar posições diferentes, independente de recursos de câmera como afastamento ou aproximação. Entretanto, uma desvantagem é que alguns resultados destoam muito da imagem modelo.

Embora a técnica de wavelet não fosse utilizada para identificação de faces, como descrito nos trabalhos relacionados (Capítulo 3), os resultados puderam identificar cenas de notícias em que dois âncoras estão presentes na imagem lado-a-lado, como pode ser observado na Figura 4.10, a qual representa o melhor resultado para esta técnica visualizado na Tabela 4.6 relacionada ao Jornal Nacional de 03/03/2010. A característica de salientar a posição espacial dos objetos demonstrou ser uma abordagem eficiente para este caso, mesmo que a imagem modelo seja de apenas um âncora. Outra vantagem desta técnica é identificar um único âncora na imagem, mas em posições diferentes (mais a esquerda ou mais a direita) e também independente de recursos de câmera, como aproximação ou afastamento da imagem do âncora.

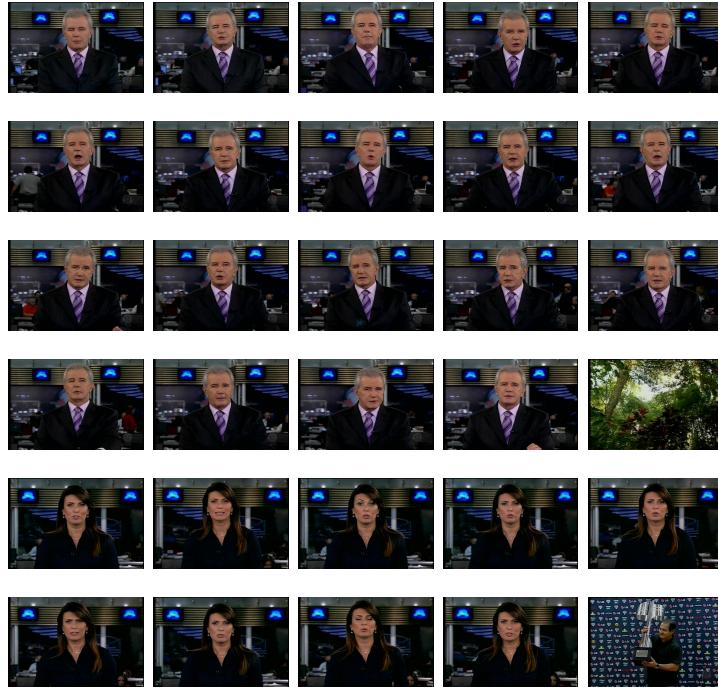


Figura 4.9: Imagens resultantes da técnica de histograma local para o Jornal da Record de 23/07/2010



Figura 4.10: Imagens resultantes da técnica de wavelet para o Jornal Nacional de 09/03/2010

4.4.5 Técnica Multimodal

A Tabela 4.7 apresenta os resultados de todos os telejornais para as m melhores cenas rankiadas, onde m é igual a 30 e 15.

Tabela 4.7: Resultados da técnica multimodal

	Técnica Multimodal			
	30		15	
	P	R	P	R
JN 22-02	60.0	55.6	66.7	31.1
JN 02-03	86.7	42.9	86.7	23.8
JN 03-03	80.0	60.5	100	30.2
JN 04-03	86.7	53.3	83.3	33.3
JN 09-03	83.3	55.3	93.3	27.7
JH 02-03	70.0	60.0	86.7	37.1
JH 03-03	60.0	40.0	86.7	28.9
JH 05-03	70.0	51.2	80.0	29.3
JH 06-03	76.7	63.9	86.7	36.1
JH 10-03	73.3	61.1	80.0	33.3
JR 16-07	70.0	36.8	86.7	22.8
JR 19-07	70.0	40.4	80.0	23.1
JR 20-07	70.0	37.5	73.3	19.6
JR 21-07	73.3	45.8	93.3	29.2
JR 23-07	76.7	41.8	73.3	20.0
JG 04-03	90.0	77.1	93.3	40.0
JG 08-09	70.0	70.0	73.3	36.7
JG 09-09	70.0	58.3	93.3	38.9
JG 21-09	63.3	57.6	86.7	39.4
JG 23-09	80.0	72.7	93.3	42.4
\bar{x}	74.0	54.1	84.8	31.1

Observa-se na Tabela 4.7 que o quanto maior o valor de m , mais eficaz é o desempenho da técnica multimodal quando comparada com as outras técnicas de imagem. Em outras palavras, a diferença do valor da média aritmética, quando comparadas com as outras técnicas de imagem, é muito maior para os valores de $m = 30$ em detrimento dos valores de $m = 15$. Para $m = 15$ a diferença é de aproximadamente 5% para o melhor resultado de precisão da melhor técnica de imagem (histograma global) em relação à multimodal. Com $m = 30$ a diferença aumenta para 17% comparando as mesmas duas técnicas, ocasionando uma melhora expressiva de desempenho com o uso da técnica multimodal.

Os resultados obtidos na Tabela 4.7 demonstram que o objetivo da técnica multimodal, o qual é salientar e incluir as vantagens de cada técnica que a compõe, foi atingido. Assim, tanto a identificação dos âncoras na imagem, quanto a detecção de silêncio, principalmente entre as propagandas comerciais, quanto o momento da fala dos âncoras, foram integrados, obtendo resultados significativos. O bom desempenho também ocorre pelo uso dos *closed-captions* e seus resultados expressivos, sendo uma metodologia ainda não adotada no desenvolvimento de técnicas multimodais para identificação de transição de cenas.

Um estudo mais detalhado comparando o desempenho das técnicas de imagens junta-

mente com a técnica multimodal, com o valor de $m=30$ (caso mais próximo de recuperação da maioria das cena), pode ser observado nas figuras a seguir, as quais contêm gráficos do tipo revocação x precisão que indicam a precisão para cada valor de revocação. Isso significa que é mostrada a performance da técnica incrementalmente para cada cena retornada. Idealmente, quanto mais a curva está próxima do topo do gráfico, melhor é o desempenho da técnica.

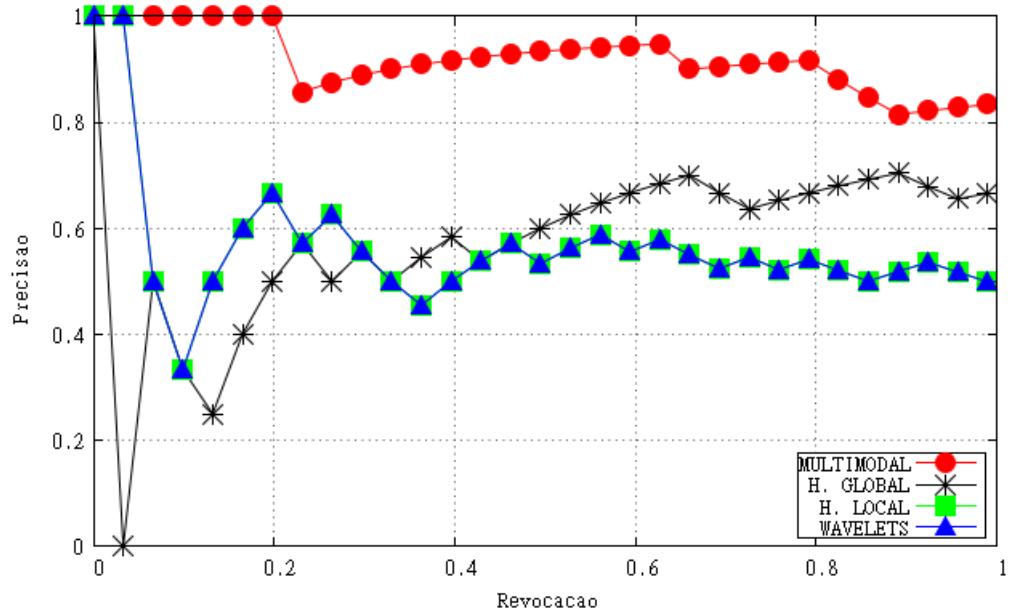


Figura 4.11: Jornal Nacional 22/02/2010

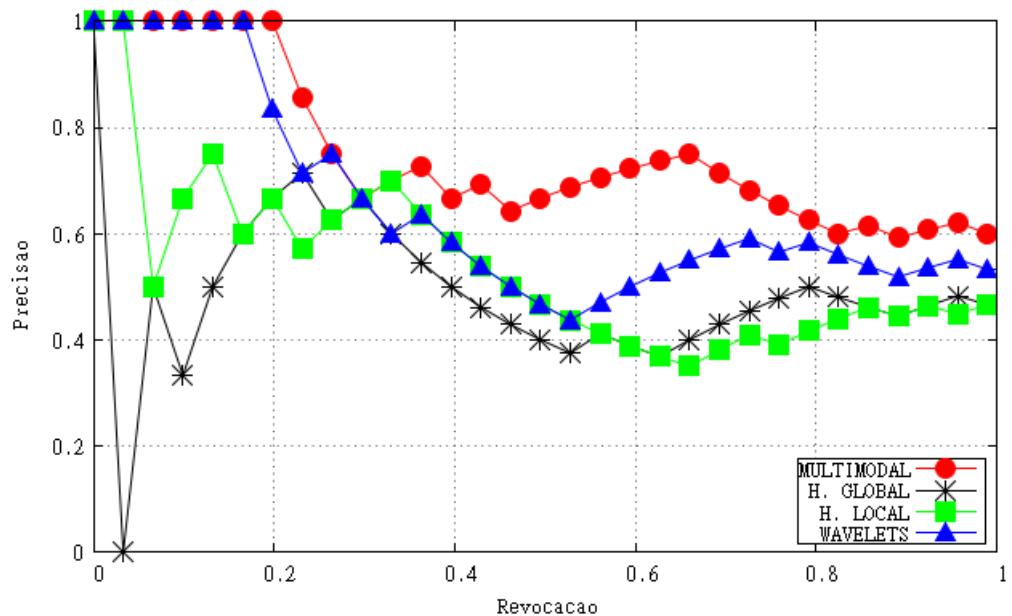


Figura 4.12: Jornal Nacional 02/03/2010

As Figuras 4.11 a 4.15 ilustram os gráficos correspondentes a cada edição capturada do telejornal Jornal Nacional. Em todas essas edições a técnica multimodal teve desempenho

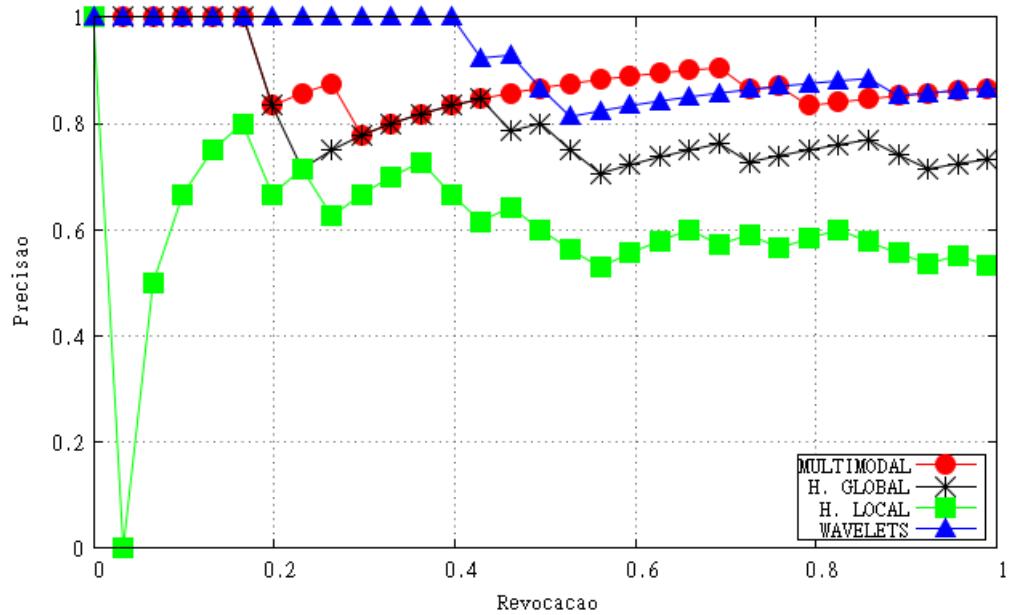


Figura 4.13: Jornal Nacional 03/03/2010

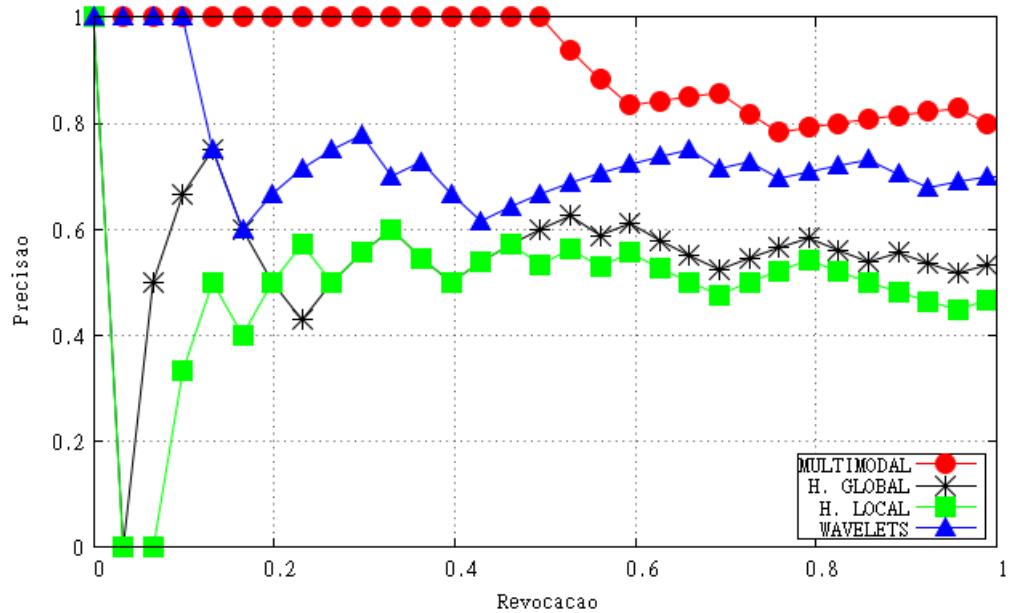


Figura 4.14: Jornal Nacional 04/03/2010

melhor, com apenas a ressalva de que na edição 03/03/2010 a técnica wavelet teve um resultado igual a da técnica multimodal.

Os resultados do telejornal Jornal Hoje são apresentados nas Figuras 4.16 a 4.20. Todas as edições apresentam resultado melhor para a técnica multimodal. O detalhe destes resultados está na edição de 10/03/2010, que tanto a técnica de histograma global quanto a wavelet obtiveram o mesmo resultado que a multimodal.

Os resultados do Jornal da Record são apresentados nas Figuras 4.21 a 4.25. Assim como nos telejornais anteriores, todas as edições apresentaram resultado melhor para

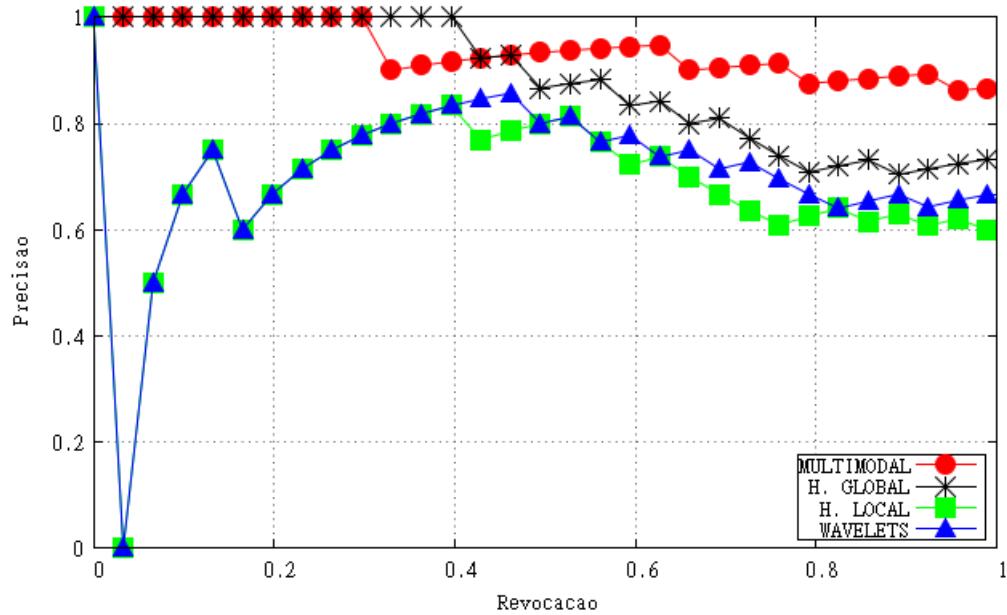


Figura 4.15: Jornal Nacional 09/03/2010

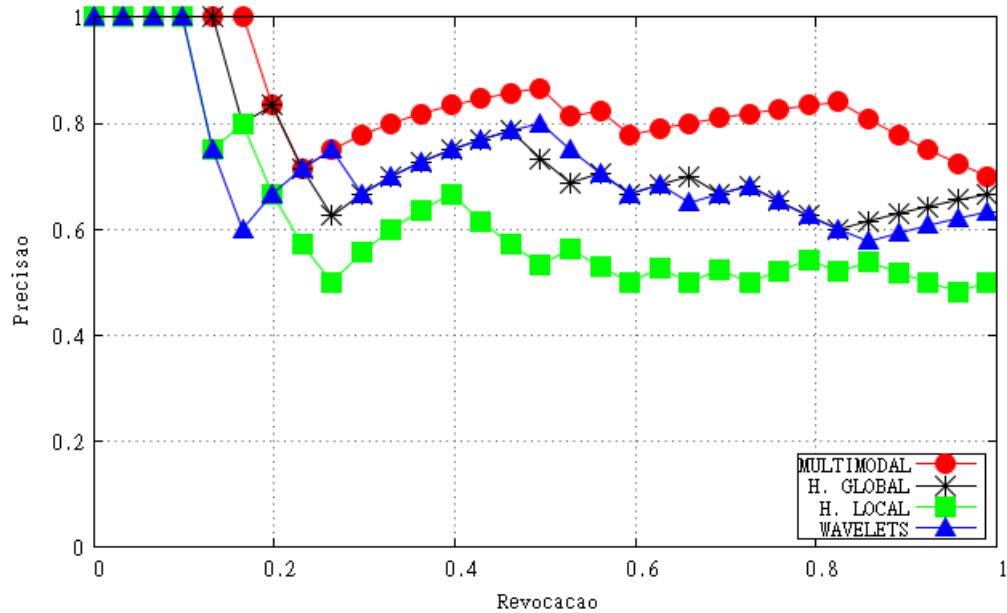


Figura 4.16: Jornal Hoje 02/03/2010

a técnica multimodal. Somente a edição de 20/07/2010 demonstra que a técnica de histograma local possui os mesmos resultados que a multimodal.

Por fim, as Figuras 4.26, 4.27, 4.28, 4.29 e 4.30 apresentam os resultados para o Jornal da Globo. Em apenas uma edição a técnica multimodal não obteve desempenho melhor, especificamente a edição de 21/09/2010, a qual a wavelet foi mais eficiente, detectando uma cena a mais.

Com base nestes resultados é possível afirmar que a técnica multimodal possui melhor desempenho e, portanto, é a mais eficiente, para todos os quatro telejornais analisados.

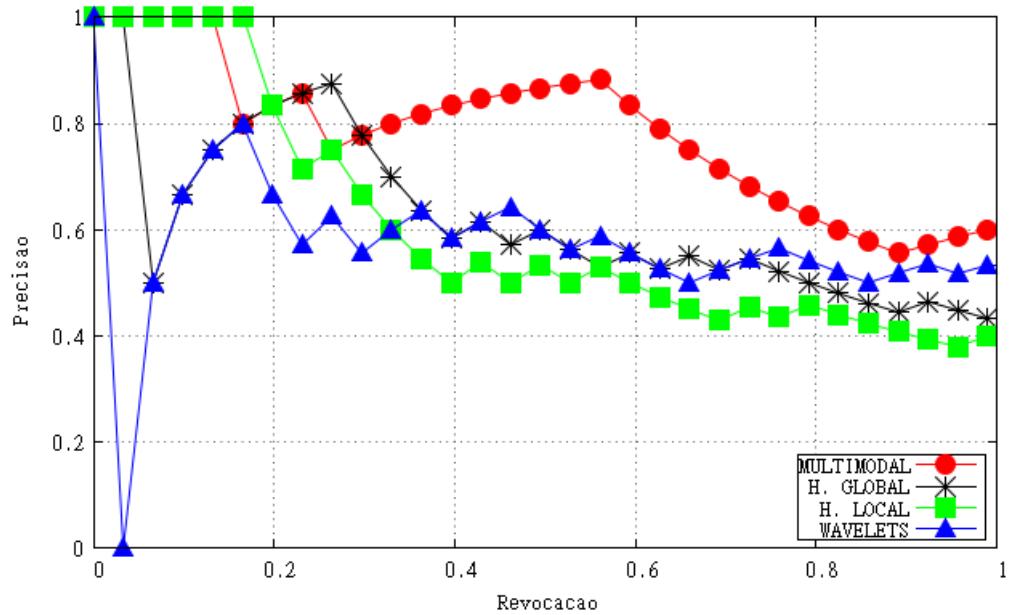


Figura 4.17: Jornal Hoje 03/03/2010

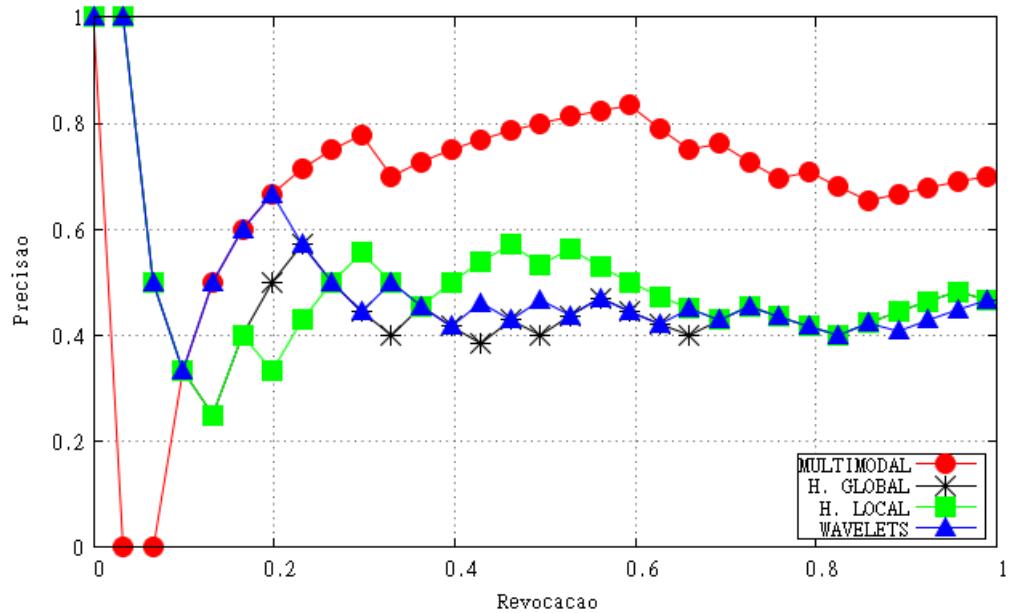


Figura 4.18: Jornal Hoje 05/03/2010

Resultados que se aproximam da técnica multimodal ocorrem porque na edição analisada, a grande parte das cenas são iniciadas com a imagem do âncora, favorecendo as técnicas que consideram posição espacial, como o histograma local e as wavelets. Essa constatação retifica a idéia de que definir o início de uma cena como a imagem do âncora é uma boa proposta para um cenário que as notícias dos telejornais são iniciadas com a imagem de um âncora, entretanto, a técnica multimodal pôde identificar cenas que não possuem esse padrão, ou seja, cenas que um âncora apresenta, mas a imagem é formada por mais de um e cenas em que são apresentadas sem a imagem de qualquer âncora.

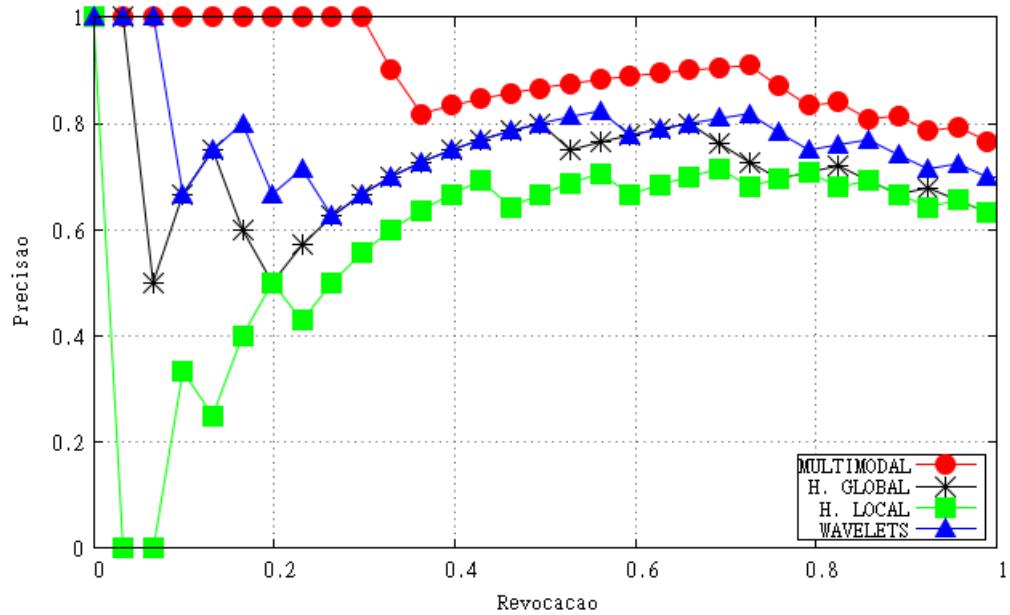


Figura 4.19: Jornal Hoje 06/03/2010

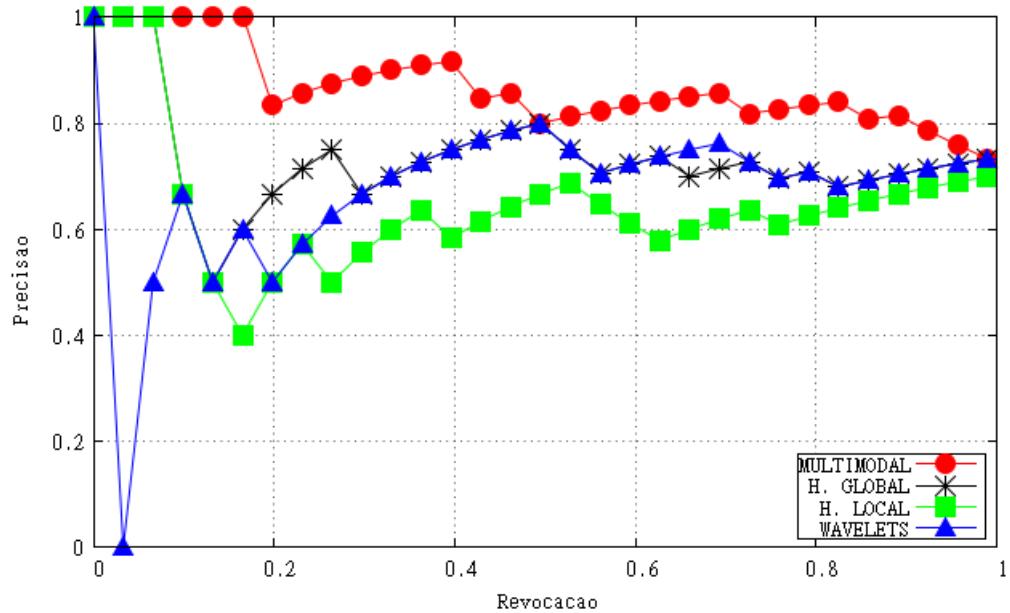


Figura 4.20: Jornal Hoje 10/03/2010

A comparação da técnica multimodal desenvolvida e apresentada neste trabalho com qualquer técnica de transição de cenas descrita na literatura é uma tarefa difícil, tendo como principais motivos:

- A principal base de dados da literatura (TRECVID) não é aberta, mas restrita aos participantes do evento (De Santo et al., 2006b).
- O modelo visual da notícia dos telejornais são diferentes e variam de emissora para emissora (De Santo et al., 2006b; Colace et al., 2005). Por exemplo, nos telejor-

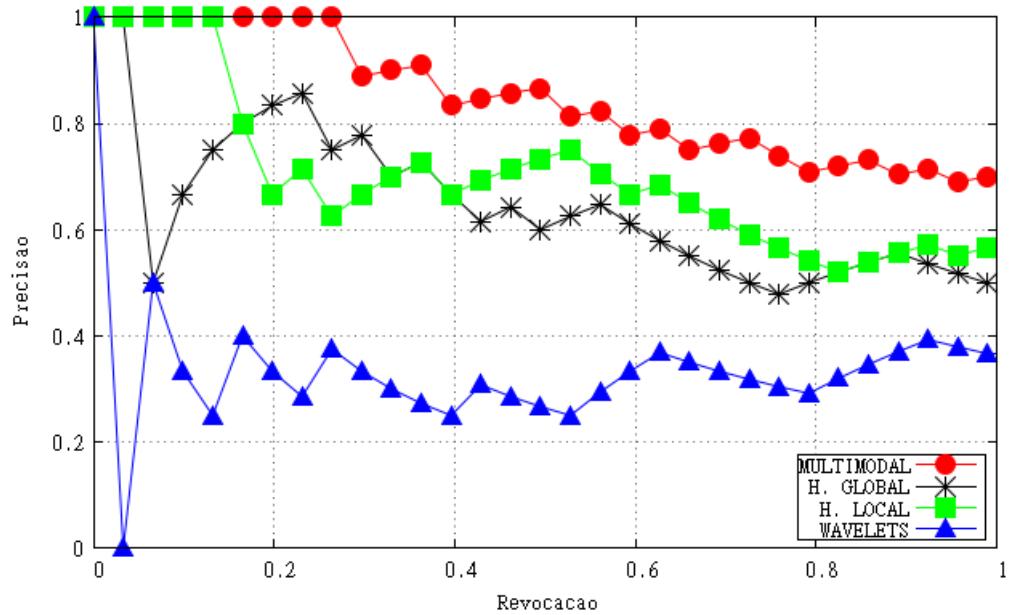


Figura 4.21: Jornal da Record 16/07/2010

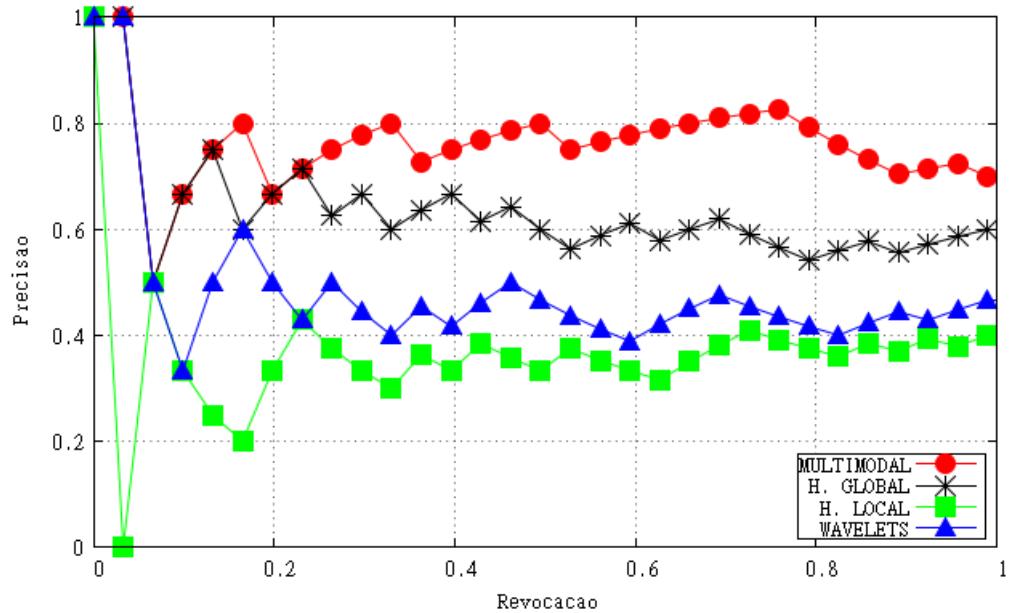


Figura 4.22: Jornal da Record 19/07/2010

nais estrangeiros CNN e ABC a primeira imagem do âncora ocorre entre 25 e 55 segundos do vídeo (Misra et al., 2010). Isso não ocorre nos telejornais brasileiros, pois a primeira tomada do âncora ocorre entre 1 e 15 segundos. Outro exemplo é a quantidade de âncoras que apresentam as notícias ou mesmo o padrão de plano de fundo do telejornal (dinâmico ou estático), o qual, geralmente varia de emissora para emissora.

- Em relação aos *closed-captions*, segundo Hauptmann & Witbrock (1998), telejornais americanos (CNN e ABC) muito usados como base para testes, possuem sintaxe di-

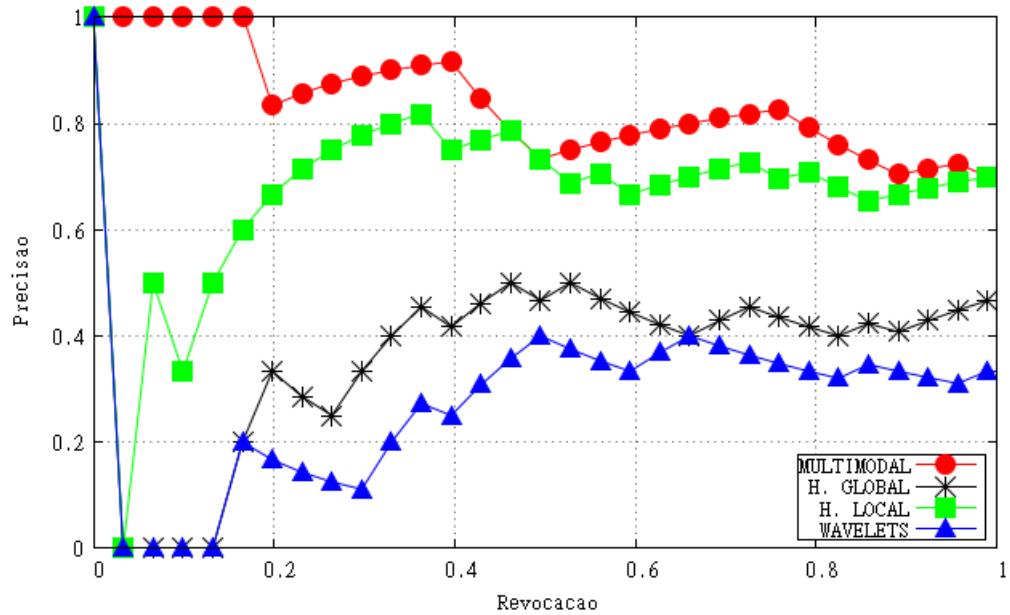


Figura 4.23: Jornal da Record 20/07/2010

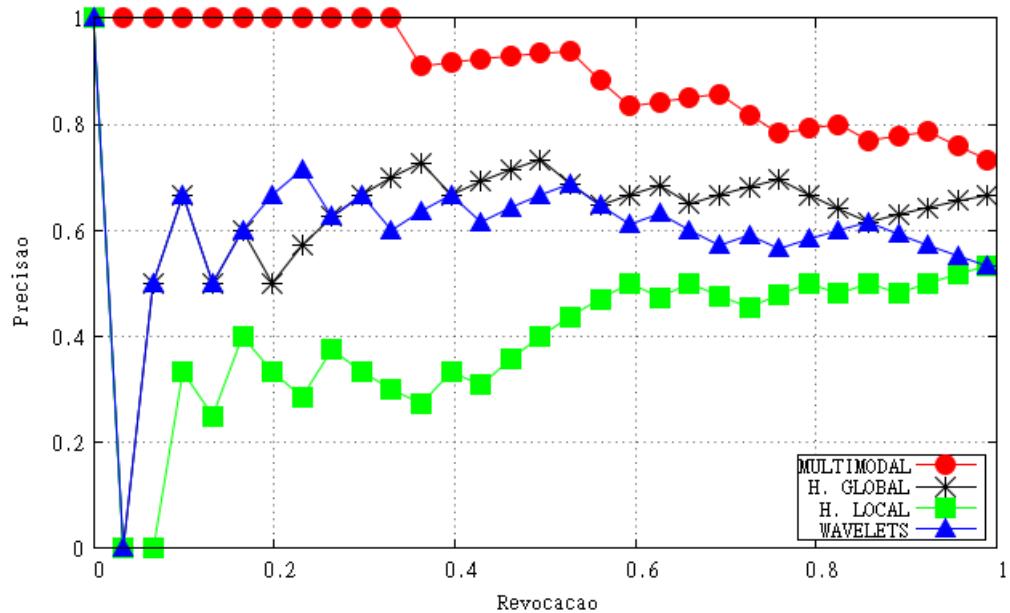


Figura 4.24: Jornal da Record 21/07/2010

ferente associados aos símbolos. Por exemplo, o símbolo “>>” indica mudança do locutor, isto é, não necessariamente indica a fala do âncora, como nos telejornais brasileiros. Ainda, “>>>” e <Locutor>: são símbolos que indicam, respectivamente, mudança de tópico e o nome do locutor, inexistentes no *closed-caption* dos telejornais brasileiros.

Uma possível solução é aplicar as técnicas da literatura usando como base os telejornais brasileiros e, por consequência, adaptar os conceitos visuais inerentes à produção destes telejornais.

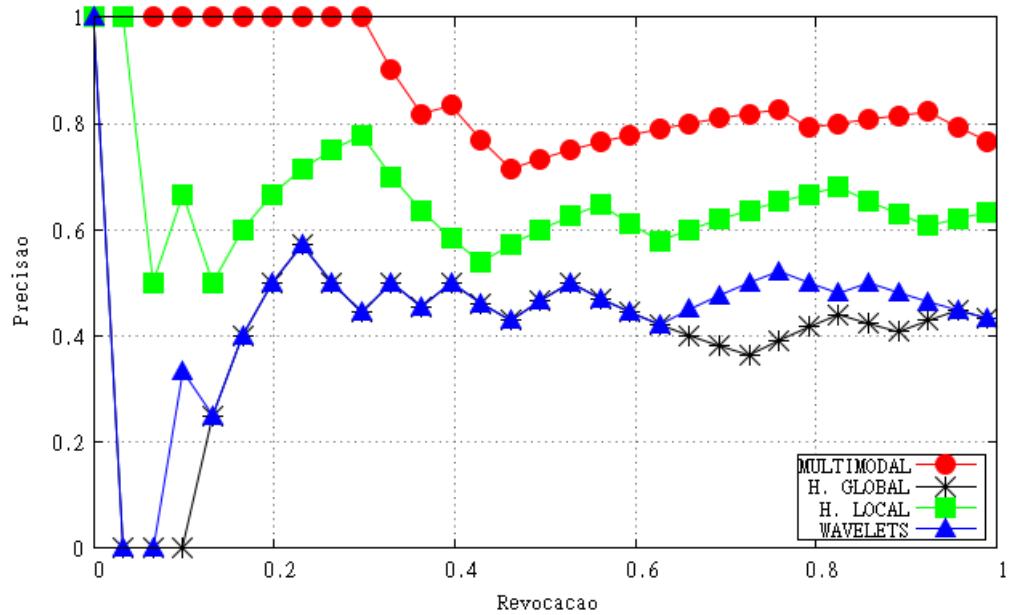


Figura 4.25: Jornal da Record 23/07/2010

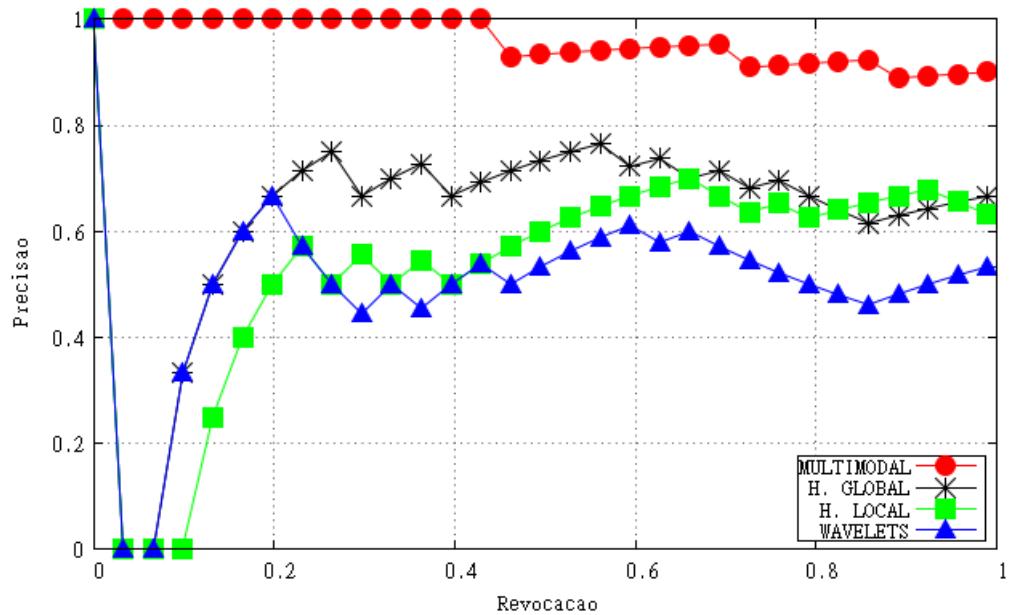


Figura 4.26: Jornal da Globo 04/03/2010

4.5 Considerações Finais

Como observado no decorrer dos capítulos deste trabalho, há uma demanda crescente na área de recuperação de informação em vídeo digital para o uso de mais de uma fonte de informação, especificamente com uso de várias mídias. Deste modo foi desenvolvida uma técnica multimodal que segmenta cenas de telejornais empregando técnicas visuais, sonoras e textuais que, utilizadas em conjunto possam melhorar a semântica associada ao conteúdo.

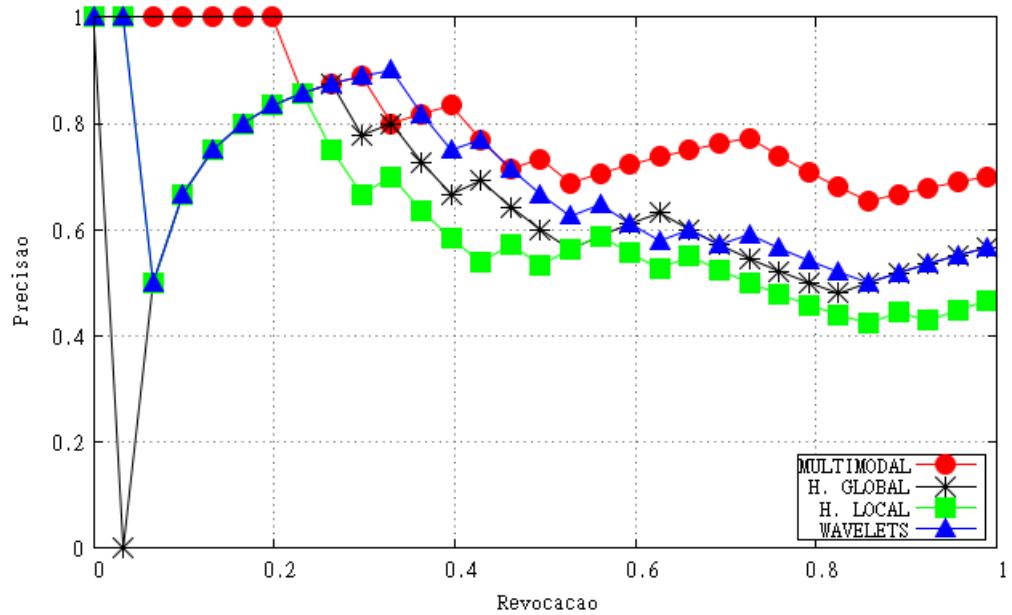


Figura 4.27: Jornal da Globo 08/09/2010

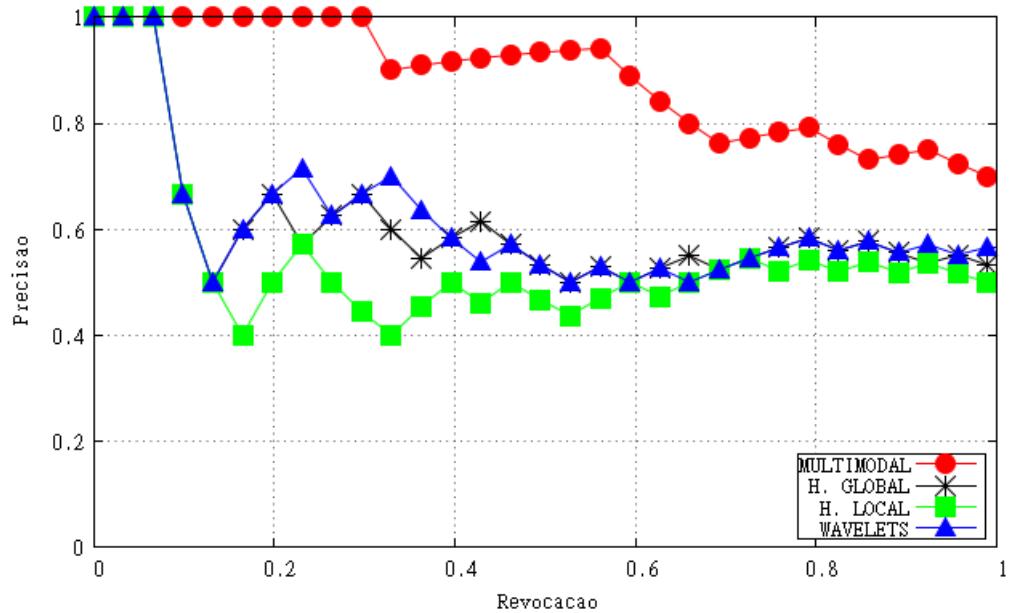


Figura 4.28: Jornal da Globo 09/09/2010

Para o desenvolvimento da técnica multimodal que integre informações de todas as mídias, foram empregadas uma técnica de texto, duas de áudio e três de imagens. A metodologia seguida foi aplicar cada técnica separadamente na base de vídeos, coletando os resultados e, posteriormente, comparando com os resultados da técnica multimodal, a qual é composta pela união das vantagens de cada técnica monomodal, usando para isso um determinado conceito de cena, os quais são necessários, mas não discutidos em muitos trabalhos da área.

Em relação às técnicas de imagem, com os resultados obtidos por intermédio do his-

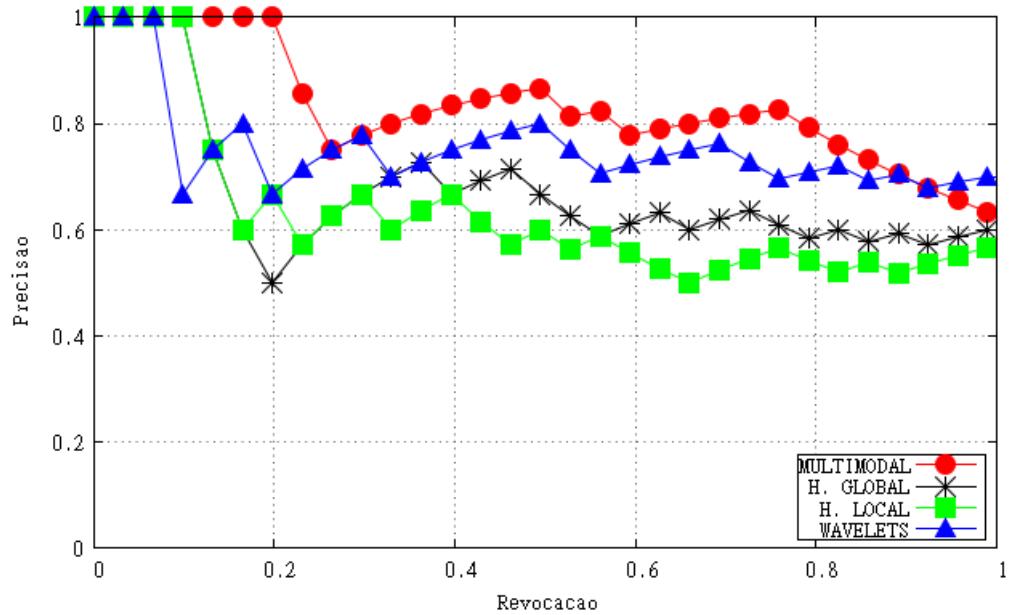


Figura 4.29: Jornal da Globo 21/09/2010

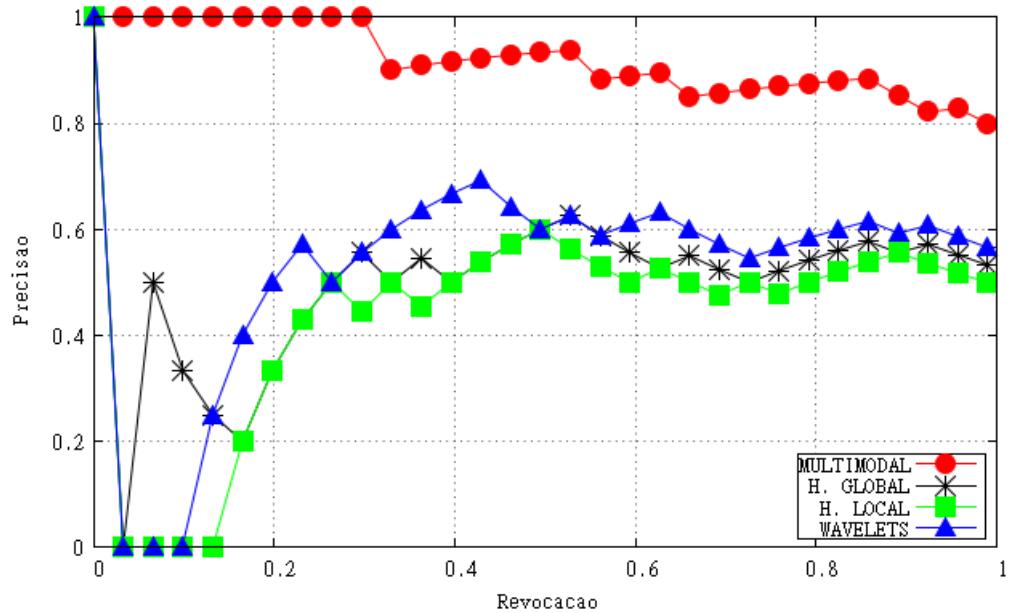


Figura 4.30: Jornal da Globo 23/09/2010

tograma global foi possível observar que a técnica mostrou ser eficiente para recuperar imagens quando aparece um único âncora, independente da quantidade de âncora no telejornal, mesmo que seja restrito a imagens com plano de fundo estático, que apresentem o âncora na mesma posição da imagem modelo e que não detectem cenas que iniciem com a imagem de dois âncoras. Uma vantagem desta técnica foi identificar cenas que representem notícias de climas em determinados telejornais, devido ao mesmo plano de fundo na imagem do âncora e do repórter que apresenta a notícia de clima.

Os resultados provenientes do histograma local mostram que não somente os resulta-

dos foram eficientes para todos os casos do histograma global, com exceção da detecção de cenas relacionadas às notícias de clima, como também na detecção de imagens de âncora nas mais diversas posições, ou seja, alguns momentos os âncoras são filmados por uma câmera mais distante ou com efeitos de aproximação e afastamento da lente da câmera, alterando a posição espacial da pessoa do âncora na imagem. Para esses casos, o histograma local demonstrou considerar essas situações em determinados telejornais. Devido às características de salientar a textura dos objetos na imagem, a técnica de wavelet foi capaz de identificar imagens com características distintas. Identificação de um ou dois âncoras na imagem, independente de posição ou recursos de câmeras fazem parte das vantagens do uso desta técnica. Uma desvantagem é que algumas vezes em determinados vídeos, um entrevistado ou um repórter pode retornar como falso positivo da técnica.

Mesmo que a identificação do âncora seja o principal indício de uma transição de cena, podem ocorrer situações em que o âncora não representa esse início e situações em que o início da cena não tem a imagem do âncora. Para contornar essas duas situações, as técnicas de áudio e texto auxiliam na segmentação de cena. O áudio, por intermédio de intervalos de silêncio, visa identificar a pausa do âncora para apresentar outra notícia, a qual representa outra cena, ou ainda intervalos comerciais. Por meio do *closed-caption* é possível identificar o exato momento em que ocorre a fala do âncora, independente se existe uma imagem associada a esta fala.

Por fim, os resultados, tanto das técnicas quando aplicadas em separadas, quanto da técnica multimodal, foram analisados e discutidos nos telejornais, contrapondo as vantagens e desvantagens de cada uma. O uso da técnica multimodal é justificado, pois melhora significativamente os resultados dos telejornais analisados se comparado às outras técnicas, principalmente quando aumenta a quantidade de cenas que se deseja recuperar.

Conclusões e Trabalhos Futuros

5.1 Considerações Iniciais

Neste trabalho foi desenvolvida uma técnica multimodal que integra informações de todas as mídias do vídeo digital. Para o desenvolvimento desta técnica foi necessário desenvolver técnicas que extraiam informações de cada mídia em separado para que pudesse agrupá-las posteriormente, obtendo desse modo mais semântica associada ao conteúdo.

A necessidade de realizar a segmentação de vídeos está diretamente relacionada à área de personalização e adaptação de conteúdo, a qual tem como objetivo proporcionar a melhor versão e a melhor maneira de disponibilizar o conteúdo multimídia para o usuário. Baseado nisso, este trabalho atuou na realização da segmentação de estruturas denominadas cenas, as quais são representadas nos telejornais como notícias, vinhetas ou comerciais, contribuindo, portanto, em uma navegação e interação mais intuitiva para o usuário.

As próximas seções deste capítulo descrevem as contribuições deste trabalho (Seção 5.2), discutem as limitações do mesmo (Seção 5.3) e, por fim, apresentam sugestões para trabalhos futuros (Seção 5.4).

5.2 Contribuições

Basicamente este trabalho de mestrado teve como principais contribuições:

- Desenvolvimento de uma técnica que tenha como característica ser mais geral na detecção de cenas de telejornais, pois os trabalhos atuais, apesar de reportar resultados e metodologias das técnicas empregadas, não definem como é a semântica associada ao segmento que analisaram (cenas).

- Uso da multimodalidade para melhorar resultados da identificação de cenas, quando comparado com os resultados obtidos das técnicas monomodais.
- Contribuição para a área de personalização e adaptação de conteúdo (P&A) por meio de uma técnica de segmentação de telejornais em cenas.
- Criação e uso de uma base de vídeos heterogênea para validar a técnica desenvolvida, contendo várias edições de telejornais brasileiros de mais de uma transmissora de TV e com datas de captura distintas.

5.3 Limitações

Em relação às limitações relacionadas ao desenvolvimento e aplicação deste trabalho podemos citar:

- A principal limitação deste trabalho está relacionada ao desempenho da técnica, uma vez que este apresentou valores de precisão e revocação menores que os trabalhos relacionados. Contudo, uma ressalva importante deve ser feita: as definições de cenas são diferentes, sendo que as definições utilizadas neste trabalho abordam um conceito mais amplo. Isto torna difícil realizar uma comparação justa com as outras abordagens.
- Por não existir uma base de vídeos aberta, foi necessário criar uma base particular. Bases particulares são comuns nessa área, porém não são padronizadas e, portanto, dificultam ainda mais a comparação entre técnicas.
- A única informação textual que a técnica multimodal suporta é o *closed-caption*. Portanto, serviços como o Teletexto ou mesmo técnicas que extraiam informação de outras mídias para formato texto, como o áudio (ASR) ou imagem (OCR), não são contempladas devido a não conformidade com as características particulares do *closed-caption*.
- Restrição da técnica multimodal ao gênero de telejornais somente. Todavia, aplicá-la em outros gêneros requer adaptações, pois esta foi desenvolvida considerando a particularidade da estrutura dos telejornais.

5.4 Trabalhos Futuros

A realização deste trabalho proporcionou diversas perspectivas para o desenvolvimento de trabalhos futuros como continuação desta pesquisa, bem como idéias para novas abordagens.

Uma idéia para contornar o problema de comparação de resultados é aplicar a técnica multimodal desenvolvida em telejornais estrangeiros utilizando a conceito de cena mais geral abordado neste trabalho, ou fazer o inverso, utilizar a técnica de um trabalho, juntamente com sua definição de cenas, na base de telejornais brasileiros utilizada neste trabalho.

Investigar o uso da técnica multimodal em outras categorias de vídeos, como documentários, esportes, filmes, etc, uma vez que algumas características de suas estruturas são semelhantes, como momentos de silêncio em esportes, os quais podem indicar um pedido de tempo, ocasionando o início de uma nova cena.

Outra possibilidade é adotar uma abordagem massiva de técnicas de identificação de transição de cenas, usando uma quantidade e variedade maior de métodos, tendo em vista que na literatura é reportado que quanto maior a quantidade de técnicas e mídias empregadas, mais heterogênea é quantidade de informação, possibilitando melhores resultados.

Uma sugestão para trabalhos na área de segmentação de vídeos é investigar o uso de informações textuais provenientes de *closed-caption*, pois não há uma padronização deste conteúdo para telejornais de países diferentes, dificultando a segmentação de cenas quando a base é composta por tais telejornais.

Referências Bibliográficas

- Adomavicius, G. e Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17:734–749.
- Al-Hames, M., Zettl, S., Wallhoff, F., Reiter, S., Schuller, B., e Rigoll, G. (2006). A two-layer graphical model for combined video shot and scene boundary detection. In *Multimedia and Expo, 2006 IEEE International Conference on*, pp. 261 –264.
- Alatan, A. A., Akansu, A. N., e Wolf, W. (2001). Multi-modal dialog scene detection using hidden markov models for content-based multimedia indexing. *Multimedia Tools and Applications*, 14(2):137–151.
- Altheide, D. (1985). *Media Power*. Beverly Hills.
- Aner-Wolf, A. e Kender, J. R. (2004). Video summaries and cross-referencing through mosaic-based representation. *Comput. Vis. Image Underst.*, 95(2):201–237.
- Athanasiadis, E. e Mitropoulos, S. (2010). A distributed platform for personalized advertising in digital interactive tv environments. *Journal of Systems and Software*, 83(8):1453 – 1469.
- Baeza-Yates, R. A. e Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- Barla, A., Odone, F., e Verri, A. (2003). Histogram intersection kernel for image classification. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, v. 3, pp. 513–516.
- Barrios, V. M. G., Mödritscher, F., e Gütl, C. (2005). Personalization versus Adaptation? A User-centred Model Approach and its Application. In *Proceedings of I-KNOW'05*, pp. 120–127.

- Bilenko, M., Basu, S., e Mooney, R. J. (2004). Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the 21st. International Conference on Machine Learning (ICML)*, pp. 81–88.
- Biolchini, J., Mian, P., Natali, A., e Travassos, G. (2005). Systematic review in software engineering: Relevance and utility. Technical report, COPPE/UFRJ.
- Boggs, J. M. e Petrie, D. W. (2000). *The Art of Watching Films*. Mayfield Publishing Company, 5th ed.
- Brezeale, D. e Cook, D. (2008). Automatic video classification: A survey of the literature. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(3):416 –430.
- Cao, J.-R. (2007). Algorithm of scene segmentation based on svm for scenery documentary. pp. 95–98.
- Cao, Y., Tavanapong, W., Li, D., Oh, J., Groen, P. C. d., e Wong, J. (2004). A visual model approach for parsing colonoscopy videos. In Enser, P., Kompatsiaris, Y., O'Connor, N. E., Smeaton, A. F., e Smeulders, A. W. M., editores, *Image and Video Retrieval*, v. 3115 of *Lecture Notes in Computer Science*, pp. 1969–1969. Springer Berlin/Heidelberg.
- Cavallaro, A., Steiger, O., e Ebrahimi, T. (2003). Semantic segmentation and description for video transcoding. *ICME '03: Proceedings of the 2003 International Conference on Multimedia and Expo (ICME '03)*, 3:597–600.
- Chaisorn, L., Chua, T.-S., e Lee, C.-H. (2003). A multi-modal approach to story segmentation for news video. *World Wide Web*, 6(2):187–208.
- Chen, H. e Li, C. (2010). A practical method for video scene segmentation. In *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, v. 9, pp. 153 –156.
- Chen, L., Rizvi, S. J., e Ozsu, M. T. (2003). Incorporating audio cues into dialog and action scene extraction. v. 5021, pp. 252–263. SPIE.
- Chen, L.-H., Lai, Y.-C., e Liao, H.-Y. M. (2008). Movie scene segmentation using background information. *Pattern Recogn.*, 41(3):1056–1065.
- Choi, Y. e Lee, J. (2010). Reliability and validity of scene unit coding in the visual content analysis. *Annual Meeting of the International Communication Association*, page 40.
- Chua, T.-S., Chang, S.-F., Chaisorn, L., e Hsu, W. (2004). Story boundary detection in large broadcast news video archives: techniques, experience and trends. In *Proceedings*

of the 12th annual ACM international conference on Multimedia, MULTIMEDIA '04, pp. 656–659, New York, NY, USA. ACM.

Coimbra, D. e Goularte, R. (2009). Identificação de cenas em vídeos digitais utilizando características audiovisuais. In *Proceedings of the XV Brazilian Symposium on Multi-media and the Web (WebMedia'09)*, v. 2, pp. 43–46.

Colace, F., Foggia, P., e Percannella, G. (2005). A probabilistic framework for tv-news stories detection and classification. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 1350 –1353.

Correia, P. e Pereira, F. (2004). Classification of video segmentation application scenarios. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(5):735–741.

D'Anna, L., Percannella, G., Sansone, C., e Vento, M. (2007). A multi-stage approach for news video segmentation based on automatic anchorperson number detection. In *Mobile Ubiquitous Computing, Systems, Services and Technologies, 2007. UBICOMM '07. International Conference on*, pp. 229 –234.

De Santo, M., Foggia, P., Sansone, C., Percannella, G., e Vento, M. (2006). An unsupervised algorithm for anchor shot detection. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, v. 2, pp. 1238 –1241.

De Santo, M., Percannella, G., Sansone, C., e Vento, M. (2006). Unsupervised news video segmentation by combined audio-video analysis. In Gunsel, B., Jain, A., Tekalp, A., e Sankur, B., editores, *Multimedia Content Representation, Classification and Security*, v. 4105 of *Lecture Notes in Computer Science*, pp. 273–281. Springer Berlin / Heidelberg.

Deb, S. e Zhang, Y. (2004). An overview of content-based image retrieval techniques. In *AINA '04: Proceedings of the 18th International Conference on Advanced Information Networking and Applications*, page 59, Washington, DC, USA. IEEE Computer Society.

Dimitrova, N., Zhang, H.-J., Shahraray, B., Sezan, I., Huang, T., e Zakhor, A. (2002). Applications of video-content analysis and retrieval. *Multimedia, IEEE*, 9(3):42–55.

Dong, A. e Li, H. (2006). Semantic segmentation of documentary video using music breaks. *Multimedia and Expo, 2006 IEEE International Conference on*, pp. 1825–1828.

Dunlop, H. (2010). Scene classification of images and video via semantic segmentation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pp. 72 –79.

Eickeler, S. e Muller, S. (1999). Content-based video indexing of tv broadcast news using hidden markov models. In *Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on*, v. 6, pp. 2997 –3000 vol.6.

- Fan, J., Gao, Y., e Luo, H. (2008). Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. *Image Processing, IEEE Transactions on*, 17(3):407–426.
- Fan, J., Gao, Y., Luo, H., e Jain, R. (2008). Mining multilevel image semantics via hierarchical classification. *IEEE Transactions on Multimedia*, 10(2):167–187.
- Fang, Y., Zhai, X., e Fan, J. (2006). News video story segmentation. In *Multi-Media Modelling Conference Proceedings, 2006 12th International*, page 4 pp.
- fei MA, Y., sheng BAI, X., you XU, G., e chun SHI, Y. (2001). Research on anchorperson detection method in news video. *Journal of Software*, 12(3):377–382.
- for Standardisation, I. O. (2002). MPEG-4 Description.
- Goularte, R. (2003). *Personalização e adaptação de conteúdo baseadas em contexto para TV Interativa*. PhD thesis, ICMC-USP, São Carlos.
- Graber, D. A. (1990). Seeing is remembering: How visuals contribute to learning from television news. *Journal of Communication*, 40:134–156.
- Gruber, T. R. (1993). A translation approach to portable ontology specifications. *KNOWLEDGE ACQUISITION*, 5:199–220.
- Gu, Z., Mei, T., Hua, X.-S., Wu, X., e Li, S. (2007). Ems: Energy minimization based video scene segmentation. *Multimedia and Expo, 2007 IEEE International Conference on*, pp. 520–523.
- Hampapur, A., Weymouth, T., e Jain, R. (1994). Digital video segmentation. In *Proceedings of the second ACM international conference on Multimedia*, MULTIMEDIA '94, pp. 357–364, New York, NY, USA. ACM.
- Hanjalic, A. (2004). *Content-Based Analysis of Digital Video*. Kluwer Academic Publishers. 193 pags.
- Hanjalic, A., Kakes, G., Lagendijk, R. L., e Biemond, J. (2001). Indexing and retrieval of tv broadcast news using dancers. *Journal of Electronic Imaging*, 10(4):871–882.
- Harb, H. e Chen, L. (2006). Audio-based description and structuring of videos. *International Journal on Digital Libraries*, 6(1):70–81.
- Hare, J. S., Lewis, P. H., Enser, P. G. B., e Sandom, C. J. (2006). Mind the gap: another look at the problem of the semantic gap in image retrieval. v. 6073, page 607309. SPIE.

- Hauptmann, A. e Witbrock, M. (1998). Story segmentation and detection of commercials in broadcast news video. In *Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on*, pp. 168 –179.
- Hoi, S. e Lyu, M. (2007). A multimodal and multilevel ranking framework for content-based video retrieval. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, v. 4, pp. IV–1225 –IV–1228.
- Houaiss, A. (2001). *Dicionário Houaiss de Língua Portuguesa*. Villar Ms, Rio de Janeiro, 1a edição ed. 572p.
- Hua-Yong, L. e Tingting, H. (2009). Content-based story segmentation of news video by multimodal analysis. In *Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on*, v. 7, pp. 423 –426.
- Huang, J., Liu, Z., Wang, Y., Chen, Y., e Wong, E. (1999). Integration of multimodal features for video scene classification based on hmm. In *Multimedia Signal Processing, 1999 IEEE 3rd Workshop on*, pp. 53 –58.
- Huang, S. N. e Zhang, Z. Y. (2010). Scene detection in videos using mutual information. *Applied Mechanics and Materials*, 34-35:920–926.
- Jiang, H., Lin, T., e Zhang, H.-J. (2000). Video segmentation with the assistance of audio content analysis. *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, 3:1507–1510 vol.3.
- Jianping, W., Tianqiang, P., e Bicheng, L. (2009). News video story segmentation based on naive bayes model. In *Natural Computation, 2009. ICNC '09. Fifth International Conference on*, v. 6, pp. 77 –81.
- Jin, W., Shi, R., e Chua, T. S. (2004). A semi-naive bayesian method incorporating clustering with pair-wise constraints for auto image annotation. In *Proceedings of the ACM Multimedia*.
- Joyce, R. e Liu, B. (2006). Temporal segmentation of video using frame and histogram space. *Multimedia, IEEE Transactions on*, 8(1):130–140.
- Judge, T. K. e Neustaedter, C. (2010). Sharing conversation and sharing life: video conferencing in the home. In *Proceedings of the 28th international conference on Human factors in computing systems, CHI '10*, pp. 655–658, New York, NY, USA. ACM.
- Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical report, Keele University and NICTA.

- Kodukula, S. e Nazvia, M. (2011). Article: Evaluation of critical success factors for telemedicine implementation. *International Journal of Computer Applications*, 12(10):29–36. Published by Foundation of Computer Science.
- Koprinska, I. e Carrato, S. (2001). Temporal video segmentation: A survey. *Signal Processing: Image Communication*, 16:477–500(24).
- Koskela, M., Sjöberg, M., e Laaksonen, J. (2009). Improving automatic video retrieval with semantic concept detection. In Salberg, A.-B., Hardeberg, J., e Jenssen, R., editores, *Image Analysis*, v. 5575 of *Lecture Notes in Computer Science*, pp. 480–489. Springer Berlin / Heidelberg.
- Lan, D.-J., Ma, Y.-F., e Zhang, H.-J. (2004). Multi-level anchorperson detection using multimodal association. *Pattern Recognition, International Conference on*, 3:890–893.
- Lee, H., Yu, J., Im, Y., Gil, J.-M., e Park, D. (2011). A unified scheme of shot boundary detection and anchor shot detection in news video story parsing. *Multimedia Tools and Applications*, 51:1127–1145.
- Lee, J.-H., Lee, G.-G., e Kim, W.-Y. (2003). Automatic video summarizing tool using mpeg-7 descriptors for personal video recorder. *IEEE Transactions on Consumer Electronics*, 49(3):742–749.
- Li, Y., Ming, W., e Kuo, C.-C. (2001). Semantic video content abstraction based on multiple cues. In *Proc. IEEE International Conference on Multimedia and Expo ICME 2001*, pp. 623–626.
- Li, Y., Narayanan, S., e Kuo, C. (2004). Content-based movie analysis and indexing based on audiovisual cues. 14(8):1073–1085.
- Liu, H., He, T., e Zhang, H. (2007). Nbr: A content-based news video browsing and retrieval system. In *Technologies for E-Learning and Digital Entertainment*, pp. 793–800.
- Liu, W., Yang, G., e huang, X. (2009). Semantic features based news stories segmentation for news retrieval. In *Wavelet Analysis and Pattern Recognition, 2009. ICWAPR 2009. International Conference on*, pp. 258 –265.
- Liu, Y., Zhang, D., Lu, G., e Ma, W. Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40:262–282.
- Liu, Z., Huang, J., e Wang, Y. (1998). Classification tv programs based on audio information using hidden markov model. In *Multimedia Signal Processing, 1998 IEEE Second Workshop on*, pp. 27 –32.

- Liu, Z., Wang, Y., e Chen, T. (1998). Audio feature extraction and analysis for scene segmentation and classification. *The Journal of VLSI Signal Processing*, 20(1):61–79.
- Lowe, D. e Hall, W. (1999). *Hypermedia & the Web*. John Wiley & Sons Ltd.
- Lu, L., Cai, R., e Hanjalic, A. (2006). Audio elements based auditory scene segmentation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2006*, v. 5, pp. 17–20.
- LUM, W. Y. e LAU, F. C. M. (2002). A Context-Aware Decision Engine for Context Adaptation. *IEEE Pervasive Computing*, 1(3):41–49.
- Magalhães, J. e Pereira, F. (2004). Using mpeg standards for multimedia customization. *Signal Processing: Image Communication*, 19:437–456.
- Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(7):674 –693.
- Manzato, M., Coimbra, D., e Goularte, R. (2010). An enhanced content selection mechanism for personalization of video news programmes. *Multimedia Systems*, pp. 1–16.
- Manzato, M. e Goularte, R. (2008). Video news classification for automatic content personalization: A genetic algorithm based approach. In *Proceedings of the XIV Webmedia*.
- Manzato, M. G., Coimbra, D. B., e Goularte, R. (2009). Multimedia content personalization based on peer-level annotation. In *European Interactive TV Conference*, v. 1, pp. 1–8. 7th European Interactive TV Conference.
- Marchionini, G., Wildemuth, B. M., e Geisler, G. (2006). The open video digital library: A möbius strip of research and practice. *J. Am. Soc. Inf. Sci. Technol.*, 57:1629–1643.
- Misiti, M., Misiti, Y., Oppenheim, G., e Poggi, J. (1996). *Matlab: wavelet toolbox user's guide*. Natick: Math Works.
- Misra, H., Hopfgartner, F., Goyal, A., Punitha, P., e Jose, J. (2010). Tv news story segmentation based on semantic coherence and content similarity. In Boll, S., Tian, Q., Zhang, L., Zhang, Z., e Chen, Y.-P., editores, *Advances in Multimedia Modeling*, v. 5916 of *Lecture Notes in Computer Science*, pp. 347–357. Springer Berlin / Heidelberg.
- Morisawa, K., Nitta, N., e Babaguchi, N. (2005). Video scene retrieval with sign sequence matching based on audio features. pp. 121–129.
- Ng, A. Y., Jordan, M. I., e Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14.

- Ngo, T.-w., Zhang, H.-j., e Pong, T.-c. (2001). Recent advances in content based video analysis. *International Journal of Image and Graphics*, 1:1–3.
- Nievergelt, Y. (1999). *Wavelets made easy*. Boston: Birkhäuser.
- Ogawa, A., Takahashi, T., Ide, I., e Murase, H. (2008). Cross-lingual retrieval of identical news events by near-duplicate video segment detection. In Satoh, S., Nack, F., e Etoh, M., editores, *Advances in Multimedia Modeling*, v. 4903 of *Lecture Notes in Computer Science*, pp. 287–296. Springer Berlin / Heidelberg.
- Oh, J. H., Wen, Q., Hwang, S., e Lee, J. (2005). Video abstraction. In Deb, S., editor, *Video Data Management and Information Retrieval*. Idea Group Publishing.
- Pai, M., McCulloch, M., Gorman, J. D., Pai, N., Enanoria, W., Kennedy, G., Tharyan, P., e Colford Jr., J. M. (2004). Clinical research methods - systematic reviews and meta-analyses: An illustrated, step-by-step guide. *The National Medical Journal of India*, 17:86–94.
- Porter, S., Mirmehdi, M., e Thomas, B. (2003). Temporal video segmentation and classification of edit effects. *Image and Vision Computing*, 21:1097–1106.
- Ren, K., Jia, Q., Sun, H., e Zhang, W. (2010). Urban scene segmentation by graphical model. In *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, v. 4, pp. 456–460.
- Riffe, D., Lacy, S., e Fico, F. G. (1998). *Analyzing media messages: using quantitative content analysis in research*. Lawrence Elbau Assoiates, Inc.
- Rui, Y., Huang, T., e Mehrotra, S. (1998). Exploring video structure beyond the shots. In *Proc. IEEE International Conference on Multimedia Computing and Systems*, pp. 237–240.
- Sethi, I. K. e Coman, I. L. (2001). Mining association rules between low-level image features and high-level concepts. In *Proceedings of the SPIE Data Mining and Knowledge Discovery*, v. 3, pp. 279–290.
- Shao, X., Xu, C., Maddage, N. C., Tian, Q., Kankanhalli, M. S., e Jin, J. S. (2006). Automatic summarization of music videos. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(2):127–148.
- Smeulders, A., Worring, M., Santini, S., Gupta, A., e Jain, R. (2000). Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380.

- Snoek, C. G. e Worring, M. (2005). Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25:5–35.
- Snoek, C. G. M., Worring, M., e Smeulders, A. W. M. (2005). Early versus late fusion in semantic video analysis. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 399–402, New York, NY, USA. ACM.
- Souza Filho, G. L. d., Leite, L. E. C., e Batista, C. E. C. F. (2007). Ginga-J: the procedural middleware for the Brazilian digital TV system. *Journal of the Brazilian Computer Society*, 12:47 – 56.
- Sural, S., Mohan, M., e Majumdar, A. K. (2005). A soft decision histogram from the hsv color space for video shot detection. In Deb, S., editor, *Video Data Management and Information Retrieval*. Idea Group Publishing.
- Swain, M. J. e Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, 7:11–32. 10.1007/BF00130487.
- Tan, Y.-P. e Lu, H. (2002). Model-based clustering and analysis of video scenes. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, v. 1, pp. 617–620.
- Town, C. P. e Sinclair, D. (2001). Content-based image retrieval using semantic visual categories. *Society for Manufacturing Engineers. Technical Report*.
- Vasconcelos, N. (2004). On the efficient evaluation of probabilistic similarity functions for image retrieval. *IEEE Transactions on Inf. Theory*, 50(7):1482–1496.
- Wang, J. e Chua, T.-S. (2003). A cinematic-based framework for scene boundary detection in video. *The Visual Computer*, 19(5):329–341.
- Wang, J., Duan, L., Liu, Q., Lu, H., e Jin, J. (2008). A multimodal scheme for program segmentation and representation in broadcast video streams. *Multimedia, IEEE Transactions on*, 10(3):393–408.
- Wang, Y., Liu, Z., e Huang, J.-C. (2000). Multimedia content analysis-using both audio and visual clues. 17(6):12–36.
- Wen, X., Huffmire, T. D., Hu, H. H., e Finkelstein, A. (1999). Wavelet-based video indexing and querying. *Multimedia Systems*, 7:350–358. 10.1007/s005300050137.
- Xie, L., Natsev, A., e Tesic, J. (2007). Dynamic multimodal fusion in video search. In *Multimedia and Expo, 2007 IEEE International Conference on*, pp. 1499 –1502.
- Yinzi, C., Yang, D., Yonglei, G., Wendong, W., Yanming, Z., e Kongqiao, W. (2010). A temporal video segmentation and summary generation method based on shots' abrupt

and gradual transition boundary detecting. In *Communication Software and Networks, 2010. ICCSN '10. Second International Conference on*, pp. 271 –275.

Yu, H., Su, B., Lu, H., e Xue, X. (2007). News video retrieval by learning multimodal semantic information. In *Proceedings of the 9th international conference on Advances in visual information systems, VISUAL'07*, pp. 403–414, Berlin, Heidelberg. Springer-Verlag.

Yu, X., Li, C., Xu, X., Yang, S., e Wan, W. (2009). Automatic scene change detection for composed speech and music sound under low snr in compressed domain. In *Wireless Mobile and Computing (CCWMC 2009), IET International Communication Conference on*, pp. 578–581.

yu Chen, M., Li, H., e Hauptmann, A. (2010). Combining motion understanding and keyframe image analysis for broadcast video information extraction. v. 7704, page 77040H. SPIE.

Zachary, J., Iyengar, S. S., e Barhen, J. (2001). Content based image retrieval and information theory: A general approach. *Journal of the American Society for Information Science and Technology*, 52:840–852.

Zeng, X., Zhang, X., Hu, W., e Li, W. (2010). Video scene segmentation using time constraint dominant-set clustering. In Boll, S., Tian, Q., Zhang, L., Zhang, Z., e Chen, Y.-P., editores, *Advances in Multimedia Modeling*, v. 5916 of *Lecture Notes in Computer Science*, pp. 637–643. Springer Berlin / Heidelberg.

Zhai, Y. e Shah, M. (2005). A general framework for temporal video scene segmentation. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, v. 2, pp. 1111–1116.

Zhai, Y. e Shah, M. (2006). Video scene segmentation using markov chain monte carlo. *Multimedia, IEEE Transactions on*, 8(4):686–697.

Zhang, D. e Lu, G. (2003). Evaluation of similarity measurement for image retrieval. In *Neural Networks and Signal Processing, 2003. Proceedings of the 2003 International Conference on*, v. 2, pp. 928 – 931.

Zhang, D., Zhou, L., Briggs, R. O., Nunamaker, J. F., e Jr. (2006). Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information & Management*, 43(1):15 – 27.

Zhang, H., Gong, Y., Smoliar, S., e Tan, S. Y. (1994). Automatic parsing of news video. In *Multimedia Computing and Systems, 1994., Proceedings of the International Conference on*, pp. 45 –54.

- Zhang, H. J. (2002). Content-based video analysis, retrieval and browsing. Technical report, Media Computing Group, Microsoft Research Asia, Beijing.
- Zhang, S. e Wang, H. (2010). Video segmentation based on acoustic analysis. In *Artificial Intelligence and Education (ICAIE), 2010 International Conference on*, pp. 1–4.
- Zhang, Y.-J., editor (2006). *Advances in Image and Video Segmentation*. IRM Press, Hershey, Pa, USA.
- Zhao, L., Yang, S.-Q., e Feng, B. (2001). Video scene detection using slide windows method based on temporal constrain shot similarity. In *Proc. IEEE International Conference on Multimedia and Expo ICME 2001*, pp. 1171–1174.
- Zhao, M., Neo, S.-Y., Goh, H.-K., e Chua, T.-S. (2006). Multi-faceted contextual model for person identification in news video. In *Multi-Media Modelling Conference Proceedings, 2006 12th International*, page 8 pp.
- Zheng, F., Li, S., Li, H., e Feng, J. (2009). Weighted block matching-based anchor shot detection with dynamic background. In Kamel, M. e Campilho, A., editores, *Image Analysis and Recognition*, v. 5627 of *Lecture Notes in Computer Science*, pp. 220–228. Springer Berlin / Heidelberg.
- Zhu, S. e Liu, Y. (2008). Automatic scene detection for advanced story retrieval. *Expert Systems with Applications*, In Press, Uncorrected Proof.
- Zhu, S. e Liu, Y. (2008). A novel scheme for video scenes segmentation and semantic representation. In *Proc. IEEE International Conference on Multimedia and Expo*, pp. 1289–1292.
- Zhu, S. e Liu, Y. (2009). Video scene segmentation and semantic representation using a novel scheme. *Multimedia Tools and Applications*, 42:183–205. 10.1007/s11042-008-0233-0.
- Zutschi, S., Wilson, C., Krishnaswamy, S., e Srinivasan, B. (2005). *Managing Multimedia Semantics*, Cap. The role of relevance feedback in managing multimedia semantics: A survey, pp. 288–304. Idea Group Inc.

Apêndice

Revisão Sistemática

Além da revisão bibliográfica tradicional, a revisão sistemática também foi empregada como metodologia de pesquisa científica. 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, essa revisão difere da tradicional em alguns pontos.

Enquanto a revisão tradicional 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, por sua vez, tem uma busca 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). Portanto, 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.

Planejamento da Revisão

O problema de pesquisa averiguado envolve o desafio associado a manipulação de arquivos multimídia, como os vídeos digitais, pois arquivos desse tipo possuem muitas informações associadas a seu conteúdo. Mesmo com o desenvolvimento de técnicas para segmentação de vídeo de modo automático e semi-automático que visam facilitar a extração de informações e de seus conteúdos, ainda faz-se necessário melhorar tal segmentação, adicionando semântica a estas técnicas, possibilitando extração de informações de modo mais eficiente e facilitando a identificação de segmentos denominados cenas. Portanto, a principal questão da pesquisa é: “Quais desafios das técnicas/modelos/métodos de segmentação de vídeo que faz uso de semântica?”

Após várias iterações refinando o resultado da pesquisa, a *string* final de busca obtida foi: “(*video semantic* “*shot clustering*” *OR* “*scene segmentation*” -*survey*) *ou* (*video AND semantic AND* (“*shot clustering*” *OR* “*scene segmentation*” -*survey*))”. Como fontes de pesquisa, foram utilizados sítios na Web que ofereçam serviços de buscas e conteúdos para trabalhos acadêmicos, tais como Google Acadêmico¹, IEEE², ACM³, Elsevier⁴ e Springer⁵.

Condução da Revisão

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* foi efetuada para não criar nenhuma pesquisa tendenciosa do ponto de vista de um grupo recluso de autores.

O procedimento para a seleção dos trabalhos foi feito de acordo com a seguinte ordem:

- A *string* de busca foi inserida no sítio Google Acadêmico (*Google Scholar*) utilizando sua ferramenta de busca avançada. Esse sítio tem como característica incluir a base

¹<http://scholar.google.com.br>

²<http://www.ieee.org> e <http://ieeexplore.ieee.org>

³portal.acm.org/dl.cfm

⁴www.elsevier.com

⁵www.springer.com

de dados dos principais periódicos acadêmicos existentes, dentre eles ACM, IEEE, Elsevier e Springer.

- Os trabalhos selecionados foram armazenados em um programa de código aberto (do inglês, *Open Source*) denominado JabRef⁶, 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 L^AT_EX .
- Efetuou-se uma pré-seleção dos trabalhos que tem como critério a análise dos títulos dos trabalhos.
- Selecionou-se pela leitura do resumo (do inglês, *abstract*) e, se necessário, da conclusão.
- Realizou-se leitura total dos artigos para extrair dados quantitativos ao tema de pesquisa.

Análise dos Resultados

Foram selecionados, com as strings de busca formalizadas durante as etapas de planejamento e condução da revisão, 308 trabalhos da área relacionada ao tema. Na Figura 5.1 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.

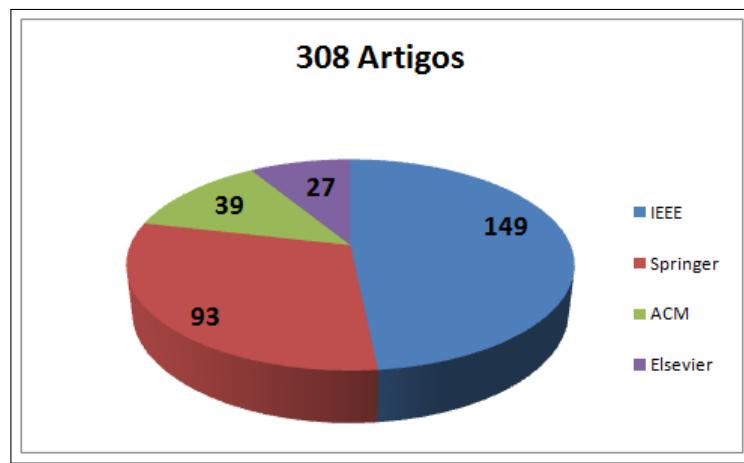


Figura 5.1: Distribuição do número de artigos por periódico

É possível observar na Figura 5.2 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.

⁶<http://jabref.sourceforge.net/>

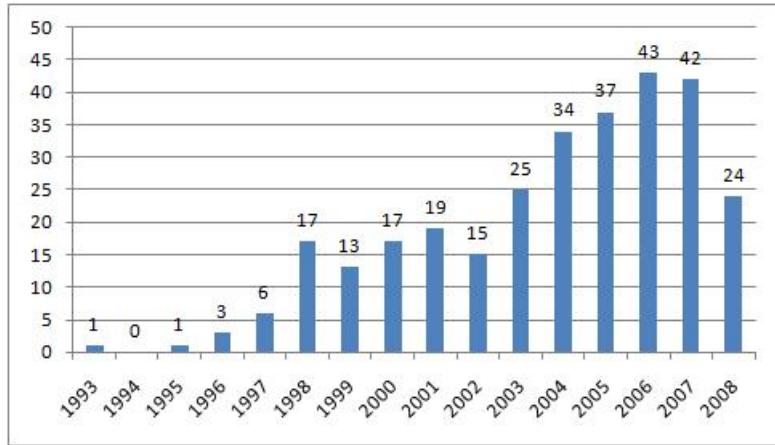


Figura 5.2: Distribuição da quantidade de artigos pelo tempo

A Figura 5.3 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 5.3: Porcentagem de produção no decorrer do tempo

A fim 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, retrieval, events, summarization, analysis. Assim, do conjunto de 227 trabalhos, 88 foram selecionados (Figura 5.4).

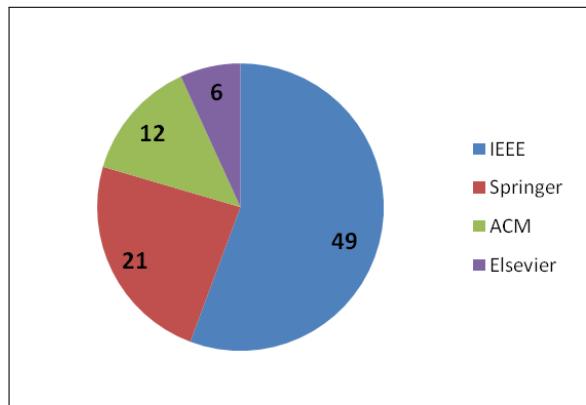


Figura 5.4: Distribuição do número de artigos após fase de seleção por resumo

Continuando a demonstração dos resultados quantitativos, a Figura 5.5 evidencia que a pesquisa é realizada em cima de um tema muito procurado pela comunidade científica dos últimos 5 anos da data da pesquisa (09/2008).

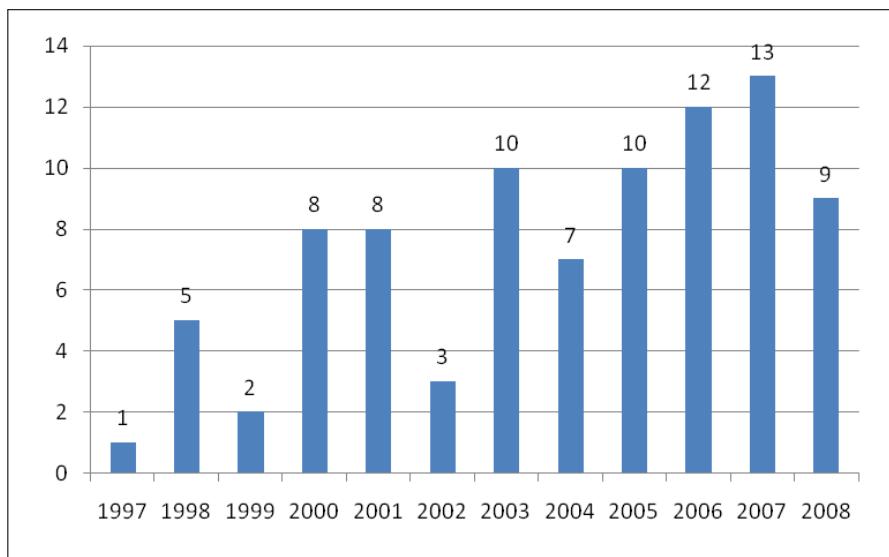


Figura 5.5: Distribuição da quantidade de artigos pelo tempo, após fase de seleção por resumo

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 5.6 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

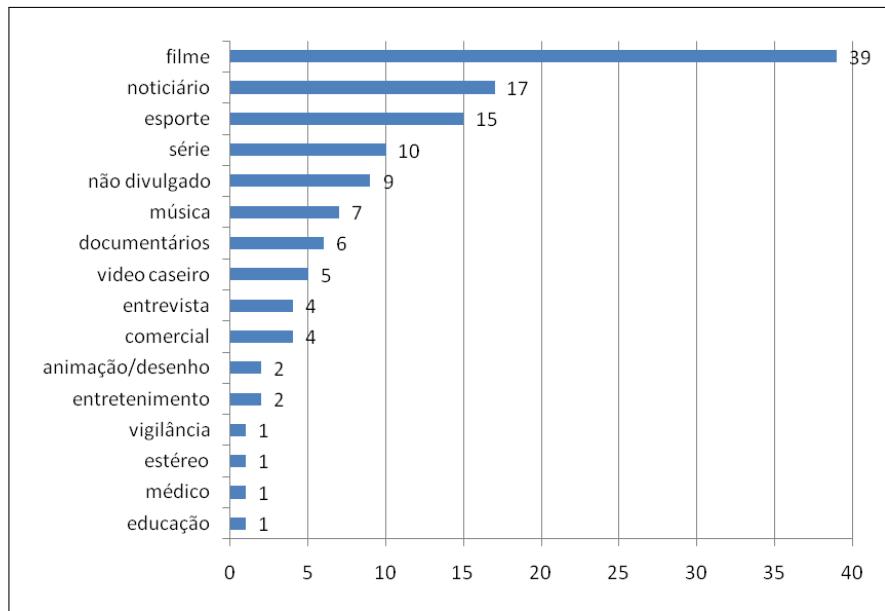


Figura 5.6: Gêneros de vídeos identificados no processo de validação das técnicas

um representa uma unidade no gráfico da Figura 5.6. 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.

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 5.7 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 destaque também as técnicas que utilizam mais de um tipo de mídia, a qual teoricamente, tem um resultado melhor, mas torna-se mais complexo.

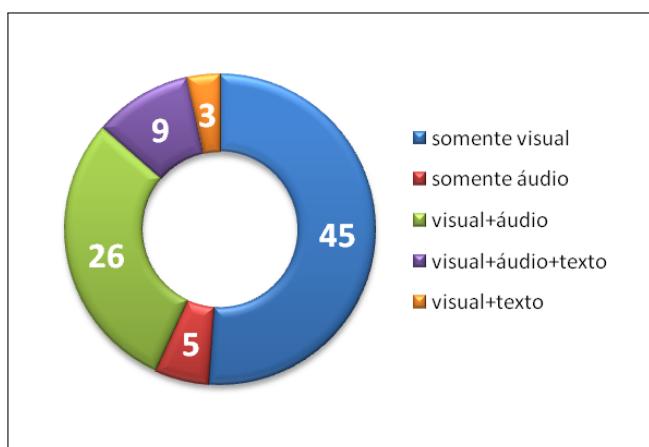


Figura 5.7: Quais tipos de mídias foram extraídas em cada trabalho

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 FlashTM⁷, 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 5.8).

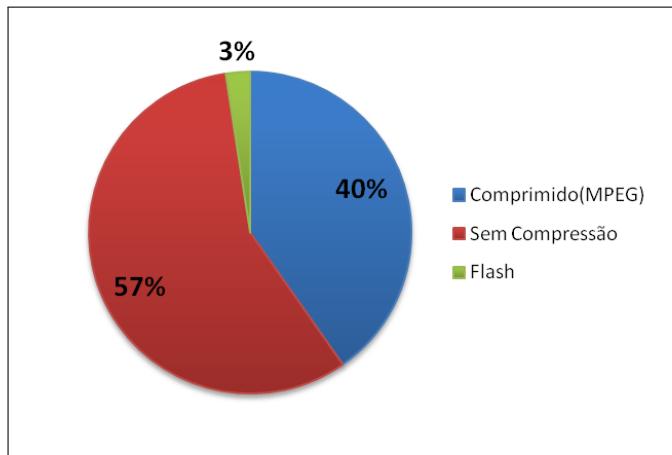


Figura 5.8: Quantidade dos trabalhos que utilizam um formato de compressão(em %)

⁷<http://www.macromedia.com/software/flash/about/>

Glossário

Adaptação – Procedimento que procura decidir a versão de conteúdo ideal para apresentação, e a melhor estratégia para gerar essa versão.

Aprendizado de Máquina – É um sub-campo da inteligência artificial dedicado ao desenvolvimento de algoritmos e técnicas que permitam ao computador aprender, isto é, que permitam ao computador aperfeiçoar seu desempenho em alguma tarefa. Enquanto que na Inteligência Artificial existem dois tipos de raciocínio: indutivo, que extrai regras e padrões de grandes conjuntos de dados, e dedutivo, a aprendizagem de máquina só se preocupa com o indutivo.

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.

ASR – *Automatic Speech Recognition*. Tecnologia criada para converter o fluxo de áudio contendo falas em texto.

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.

Closed-Caption – Sistema de transmissão de legendas que tem como objetivo permitir que os deficientes auditivos possam acompanhar os programas transmitidos. Descreve além das falas dos atores ou apresentadores qualquer outro som presente na cena, como palmas, passos, trovões, música, risos, etc.

Dispositivos Móveis – 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.

Histograma – Representação gráfica da distribuição de frequências de uma massa de medições, normalmente um gráfico de barras verticais.

HMM – Modelo Escondido de Markov (do Inglês, *Hidden Markov Model*). Modelo estatístico em que o sistema modelado é assumido como um processo de Markov com parâmetros desconhecidos, e o desafio é determinar os parâmetros ocultos a partir dos parâmetros observáveis. Os parâmetros extraídos do modelo podem então ser usados para realizar novas análises, por exemplo para aplicações de reconhecimento de padrões.

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.

Lacuna Semântica – A falta de coincidência entre as informações que se pode extrair do fluxo audiovisual e a interpretação que os mesmos dados geram para um determinado usuário em dada situação.

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.

LSA – Análise Semântica Latente (do inglês, *Latent Semantic Analysis - LSA*). É uma técnica de processamento de linguagem natural, em particular de vetores semânticos, de análise de relações entre um conjunto de documentos e os termos que eles contêm pela produção de um conjunto de conceitos relacionados aos documentos e termos.

LSI – Indexação Semântica Latente (do inglês, *Latent Semantic Indexing*). É um método

de indexação e busca que usa a SVD para identificar padrões nos relacionamentos entre os termos e conceitos contidos em uma coleção não-estruturada textual.

MDCT – A *Modified Discrete Cosine Transformation* é uma transformada de Fourier que foi desenvolvida sobre o princípio da técnica nomeada como *Time-Domain Aliasing Cancellation* (TDAC) ou Cancelamento de *Aliasing* no Domínio Temporal. A TDAC é uma técnica que torna possível reconstruir as amostras após serem aplicadas à função, de forma a não haver perdas e nem modificações durante reconstrução destas amostras, ou seja, é possível obter os valores iniciais mesmos após serem aplicados à função.

Metadado – Informações que descrevem um conteúdo, podendo criar uma indexação para uso em procedimentos de recuperação de informações.

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

OCR – Reconhecimento Ótico de Símbolos (do inglês, *Optical Character Recognition*). Tecnologia criada para reconhecer caracteres a partir de um arquivo de imagem.

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.

Perfil – O perfil é constituído por todas as informações que estão disponíveis sobre um usuário, podendo ser usadas para personalizar informações ou serviços.

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.

Segmentação – É o processo de identificar porções distintas de um documento, tais como cabeçalho, seções, parágrafos e figuras. Em multimídia, porções podem ser sequências,

cenas e quadros, por exemplo.

SNR – *Signal-Noise Ratio* (Razão Sinal-Ruído). Utilizada para medir o quanto de “ruído” (imagem granulada) uma imagem de vídeo contém, normalmente expressa em decibéis (dB). Essa medição é calculada por meio do valor da voltagem máxima atingida pelo sinal dividido pelo valor residual da voltagem que permanece quando o sinal é removido – ou seja, a quantidade de ruído no mesmo.

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.

Tomada – Nas modalidades visual e auditiva, representa uma gravação contínua ou não-interrupta de uma câmera ou microfone. Na modalidade textual, representa uma expressão textual contínua ou não-interrupta que pode estar em um nível de palavras ou sentenças.

Transformada Discreta do Cosseno – Técnica utilizada na eliminação da redundância espacial de uma imagem. Ela transforma os dados do domínio espacial para o domínio de freqüências, onde quanto mais alta é uma freqüência, menos perceptível ela é para o olho humano. Após a transformação, os dados são submetidos à quantização, que é onde a eliminação da redundância é realizada.

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.

Vídeo Digital – Esse tipo de formato, ao contrário dos vídeos analógicos, digitalizam as imagens por meio de circuitos denominados CCDs, que convertem a luz em dados comprehensíveis ao computador, digitalizando as imagens. A principal vantagem de se utilizar vídeo digital ao invés de vídeo analógico é que o primeiro contém melhor resolução do que o segundo e, além disso, o vídeo digital é reproduzido utilizando-se o sinal componente,

ao contrário dos vídeos analógicos, que utilizam o sinal S-vídeo e o sinal composto.

Wavelet – Em compressão de imagens, é utilizada para eliminar a redundância espacial. A imagem é dividida por meio de filtros, em duas componentes: uma de baixas frequências, contendo as principais informações sobre a imagem; e outra de altas frequências, contendo as informações irrelevantes. A aplicação é feita por meio da convolução do sinal com as componentes de baixa e alta frequência.

Zero-Crossing Rate – É uma taxa que identifica mudança de sinal ao longo do próprio sinal. Característica muito usada em reconhecimento de fala e recuperação de informação musical.