Segmentação multimodal de cenas em telejornais

Danilo Barbosa Coimbra
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
Av. Trabalhador São Carlense, 400
Caixa Postal 668 - 13560-970
São Carlos, SP - Brasil
danilobc@icmc.usp.br

Rudinei Goularte
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
Av. Trabalhador São Carlense, 400
Caixa Postal 668 - 13560-970
São Carlos, SP - Brasil
rudinei@icmc.usp.br

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.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval models, Information filtering*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Video*

General Terms

Design

Keywords

scene segmentation, news identification, multimodal technique, multimedia retrieval.

1. INTRODUÇÃO

No desenvolvimento de aplicações multimídia, um problema atual e recorrente é a dificuldade de acesso aos dados devido ao grande volume de dados aliado às restrições da infraestrutura de entrega e de recepção (limitação de banda, pouco memória, área limitada de visualização, por exemplo). 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 [13].

A personalização estuda meios de customizar e/ou filtrar os dados segundo as preferências, interesses e necessidades de um usuário específico. Em geral, os sistemas de personalização apresentam a necessidade de conhecer as informações contidas no conteúdo. Também chamadas de metadados, essas informações tem a função de descrever a mídia em si, como tipo de compressão, 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, dificultando a compreensão das informações, ocasionando a lacuna semântica [14]. Tal problema é caracterizado pela pouca ligação entre as informações de baixo nível obtidas com esses metadados (histogramas, identificação de pessoas, etc) e a interpretação do usuário para esse mesmo conteúdo.

Para obter informações semânticas no contexto de vídeo digital, em geral, como uma etapa de pré-processamento, realiza-se a segmentação do vídeo [6]. Em particular, a segmentação semântica, também conhecida como segmentação de cenas, é ainda um tema de pesquisa em aberto, pois contêm desafios relacionados a lacuna semântica, carecendo de investigação em técnicas e métodos que possam ser aplicados de modo a contribuir com a área de P&A.

No caso da segmentação de cenas para o gênero telejornal (foco deste trabalho), a literatura reporta que a abordagem multimodal oferece melhores resultados do que a abordagem monomodal [6]. Alguns trabalhos que fazem uso de várias mídias e exploram: a) reconhecimento de faces e histogramas para a identificação de âncoras; b) análise do áudio associado que indique pausa e mudança de assunto/locutor; c) uso de informações textuais para detecção de mudança de cena com *closed-captions*, reconhecimento de fala e reconhecimento de caracteres. Todavia, esses trabalhos utilizam definições particulares de cenas que dificultam a correta segmentação de elementos que compõem a estrutura dos telejornais. Em geral, utiliza-se a aparição do âncora (apresentador do telejornal) na imagem para determinar o início da cena. Contudo, cenas onde mais de um âncora aparece não são detectadas satisfatoriamente. Reconhecimento de faces tem sido usado para resolver esse problema. Porém, a

WebMedia'11: Proceedings of the 17th Brazilian Symposium on Multimedia and the Web. Full Papers.
October 3 -6, 2011, Florianópolis, SC, Brazil.
ISSN 2175-9642.
SBC - Brazilian Computer Society

maioria das técnicas multimodais analisadas nos trabalhos relacionados não conseguem segmentar cenas corretamente em certos trechos de telejornal, por exemplo, onde nenhum âncora aparece ou cenas compostas por assuntos distintos em uma mesma categoria de notícias, também chamados de segmentos semanticamente complexos.

Dessa maneira, o presente trabalho propõe o desenvolvimento de uma técnica multimodal que faça a segmentação de cenas em telejornais empregando técnicas visuais, sonoras e textuais, tendo como objetivo identificar as transições de cenas de acordo com segmentos semanticamente complexos, 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, juntamente com uma base de telejornais compostas por quatro noticiários, cada qual com cinco edições, obtidos de duas emissoras de TV distintas, totalizando 20 telejornais brasileiros. Os resultados obtidos demonstraram que a técnica desenvolvida tem como característica ser mais geral, considerando os trabalhos existentes e que a integração de técnicas que extraem informações de uma única mídia proporcionam resultados melhores quando comparadas com os resultados das mesmas técnicas aplicadas em separado.

O restante do artigo está organizado como segue. Na Seção 2 apresentam-se os principais conceitos relacionados a análise de vídeo digital. Alguns trabalhos relacionados ao estudo são apresentados na Seção 3. A metodologia é apresentada na Seção 4. Na Seção 5 apresenta-se a técnica desenvolvida e seus resultados são apresentados na Seção 6. Por fim, na Seção 7 apresentam-se as conclusões e trabalhos futuros.

2. CONCEITOS, DEFINIÇÕES E MÉTODOS

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

Também chamada de representação hierárquica [9], 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 1). 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.

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. As cenas são definidas como um grupo de tomadas com conteúdo correlacionado [18], também chamadas de unidades semânticas. A última

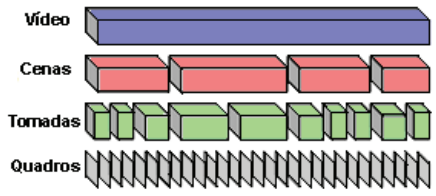

Figura 1: Estrutura do fluxo de vídeo digital

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 (*e.g.* telejornal, evento esportivo, filmes, etc).

2.1 Detecção de Cenas e a Lacuna Semântica

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 [17]. 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 vista como uma questão em aberto, também conhecida como lacuna semântica.

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

2.2 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. Logo, o conceito de cena empregado neste trabalho segue a mesma linha de Choi & Lee [3] 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), as transições de cenas acontecem quando ocorrem: *i)* transição de notícias, *ii)* transição de vinheta para notícia (ou vice-versa) ou, *iii)* transição de vinheta para comercial (ou vice-versa).

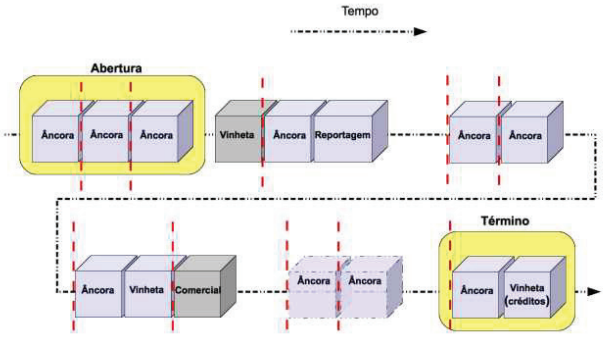

Figura 2: Composição temporal dos telejornais

De acordo com Chaisorn et al. [2], 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 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 na vertical). O detalhe nessa figura ocorre por conta dos blocos transparente de âncoras, pois representa as notícias em que não ocorrem as imagens dos âncoras, somente a fala desses apresentadores, transições também considerada no escopo deste trabalho.

No contexto relacionado à notícia, ocorre somente a peculiaridade do segmento semanticamente complexo 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 nesse trabalho considera que esse cenário é composto por três cenas, exatamente porque apesar da categoria (economia) ser a mesma, os assuntos a respeito da inflação são distintos (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 computada, 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); *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 comercias são os três tipos de segmentos que compõem o gênero telejornal.

3. TRABALHOS RELACIONADOS

Os trabalhos que envolvem a segmentação de cenas empregam o uso de uma ou mais mídias, entretanto, nesta seção, serão abordados somente os trabalhos multimodais que utilizam as três mídias do vídeo: imagem, som e texto.

Hua-Young & Tingting [7] compararam os resultados das técnicas de mídia aplicadas em separado com a técnica multimodal e conseguiram uma melhora de cerca de 11% na precisão e 5% na revocação usando a multimodalidade. Os estudos foram baseados em técnicas de OCR (do inglês, *Optical Caption Recognition*) para informação textual, comparação de histogramas para visual e detecção de clipes 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. [10], 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 [19] e também telejornais de países diferentes [8], gerando técnicas mais abrangentes. Zhao et al. [19] 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 (do inglês, *Automatic Speech Recognition*) 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. [8] 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. [11] 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. [5] 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. [16] 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 aúdio 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 que não tratam adequadamente os segmentos semanticamente complexos encontrados entre as notícias, além de não considerarem cenas que envolvam vinhetas e comerciais. Desse modo, algumas perguntas ficam em aberto: *Na abertura do telejornal, as chamadas de notícias são considerada como cenas?*; *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 indica as falas dos âncoras e o momento exato que isso ocorre.

Observou-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 utilizados, Chua et al. [4] 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.

4. TÉCNICAS EMPREGADAS

Para a composição da técnica multimodal proposta foram empregadas ao todo sete técnicas monomodais 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 e uma na parte textual. Para integrar essas técnicas foi desenvolvido um método de *ranking* (Seção 5).

A seguir serão descritas as metodologias associadas às técnicas monomodais 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* representam essa linha temporal, tornando-se indispensáveis na integração de multimodalidades, ou seja, na sincronização e alinhamento dos diferentes tipos de mídias [15].

4.1 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 representativo.
3. Aplicação de uma técnica de extração de característica visual na base de imagens. Três técnicas foram utilizadas neste trabalho: histograma de cor global e local (com blocos de 16x16 pixels) e wavelet de Daubechies com quatro coeficientes.
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 3). Para os histogramas utilizou-se a medida de similaridade denominada intersecção de histogramas e para a wavelet utilizou-se a distância euclidiana. 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.
5. Extração dos *timestamps* dos m quadros/imagens mais semelhantes, por meio do tempo (em segundos) que a imagem foi extraía do vídeo. Isso é possível porque na etapa de extração dos quadros a cada segundo, a imagem gerada foi nomeada usando o tempo da captura.

Figura 3: Imagem modelo do âncora para busca

4.2 Extração de Informação em Áudio

Foram desenvolvidas duas metodologias de extração de informação de áudio para detectar transições de cenas. Uma com a ferramenta de prpcessamento de áudio Audacity[1] e um algoritmo que calcule a raiz da média dos quadrados (do inglês, *Root Mean Square*-RMS). Ambas visam a captura de momentos de silêncio e ajudam na captura de segmentos semanticamente complexos no sentido em que identificam cenas no silêncio que existe ao mudar de assunto, sem depender do âncora.

A ferramenta Audacity encontra momentos de silêncio no vídeo usando 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:

$$A(c_i) = \begin{cases} \text{transição de cena: } r_i \leq R \\ \text{sem transição: caso contrário} \end{cases} \quad (1)$$

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 $i = 1, ..., n \in \mathbb{N}$. 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.

Com o algoritmo de RMS, foi desenvolvido um sistema em Java, com o auxílio da biblioteca nativa *javax.sound*, que revela a variação temporal da magnitude de um sinal em relação à distribuição do volume nos clipes de áudio. O cálculo é feito por um conjunto de quadros do som e computando a raiz quadrada da soma dos quadrados dos valores das amostras desses quadros, representado por [12]:

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

onde N é o tamanho do quadro, s_i é a i-ésima amostra no n-ésimo quadro (i e $n \in \mathbb{N}$). Assim, dividiu-se o fluxo de áudio em 1 quadro/s e para cada quadro obteve-se 48000 amostras. O cálculo do RMS é aplicado em cada quadro (nas 48.000 amostras) e, caso seja menor que um limiar (empiricamente: 0.5), é detectado silêncio.

4.3 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 [1], 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 repórter produtor da notícia.

Foram desenvolvidos, em linguagem Java, dois algoritmos a fim de obter as informações de transições de cenas por meio das falas dos âncoras do telejornal. Enquanto o primeiro algoritmo visa retirar duplicidades existentes nesse formato, ocasionando textos sem repetição de falas, o segundo visa identificar o momento em que ocorrem as falas dos âncoras, as quais representam os inícios das 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 é dos repórteres não aparece o nome do mesmo, somente a palavra “repórter” (*e.g.:* “>> repórter: O sistema financeiro...”). Portanto, esse algoritmo faz a análise do texto procurando por linhas que não possuam falas dos repórteres, somente as falas dos âncoras. Mesmo que seja uma técnica associada ao âncora, como a de imagem, ela auxilia na identificação de segmentos semanticamente complexos uma vez que o início de uma cena pode ocorrer somente com a fala (sem a imagem) do âncora.

5. INTEGRAÇÃO DAS TÉCNICAS

De modo geral, a técnica multimodal desenvolvida faz uso de uma tabela de espalhameto (do inglês, *hashing table*) que possui seus valores baseados em um *ranking* associado a pesos predefinidos para cada uma das técnicas. Deste modo, uma cena pode ser detectada por mais de uma técnica e as possíveis limitações de uma determinada técnica monomodal são compensadas com os pontos fortes de outra técnica, implicando em mais cenas detectadas.

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

Definição 1. *Considera-se $C = c_1, c_2, ..., c_n$ um conjunto de chaves, onde c_i é um identificador para uma transição de cena, no caso adotou-se o timestamp, com $i = 1, ..., n \in \mathbb{N}$. Como não há 2 cenas em um mesmo instante de tempo, a chave é única. Considera-se também $V = v_1, v_2, ..., v_n$ 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 é formado pela soma de pesos, os quais variam de técnica para técnica. Os valores dos pesos para cada técnica monomodal foram obtidos aplicando a mesma na base e calculando sua eficiência por meio da sua precisão e revocação (detalhes na Seção 6), ou seja, quanto maior a precisão e revocação da técnica monomodal, maior seu peso. Assim, a proposta de integração neste trabalho é definida como:

[1]http://audacity.sourceforge.net/

$$rank(v_i) = \begin{cases} \text{se} \quad T_q \in T_{cc\_texto} \\ \qquad valor(v_i) = valor(v_i) + 0.3 \\ \text{se} \quad T_q \in T_{audacity\_audio} \\ \qquad valor(v_i) = valor(v_i) + 0.1 \\ \text{se} \quad T_q \in T_{rms\_audio} \\ \qquad valor(v_i) = valor(v_i) + 0.1 \\ \text{se} \quad T_q \in T_{hglobal\_img} \\ \qquad valor(v_i) = valor(v_i) + 0.2 \\ \text{se} \quad T_q \in T_{hlocal\_img} \\ \qquad valor(v_i) = valor(v_i) + 0.1 \\ \text{se} \quad T_q \in T_{wavelet\_img} \\ \qquad valor(v_i) = valor(v_i) + 0.2 \end{cases} \tag{3}$$

onde $rank_{comb}(v_i) \in V$, com $i = 1, ..., n$; T_q é o *timestamp* a ser analisado. $T_{cc\_texto|audacity\_audio|rms\_audio|hglobal\_img|...}$ correspondem aos *timestamps* de suas respectivas técnicas.

6. RESULTADOS OBTIDOS

Como as bases de vídeos padrões são fechadas (pagas) ou não possuem os segmentos completo dos telejornais, foi necessário criar uma base para aplicar as técnicas mencionadas. A base é composta por vinte telejornais brasileiros, sendo três noticiários da Rede Globo de Televisão e um noticiário da Rede Record de Televisão, cada noticiário contendo cinco edições. Os vídeos foram capturados usando a codificação MPEG-2, com resolução de imagem de 720 x 480 pixels, a 30 quadros por segundo e com áudio a 48 kHz de frequência. Informações mais detalhadas da estrutura temporal dos telejornais são apresentadas na Tabela 1.

Tabela 1: Informações da estrutura temporal dos telejornais

Telejornal	Tempo	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

A abordagem para a execução dos testes e coleta dos resultados foi efetuada segmentando cada vídeo da base manualmente, anotando os *timestamps* de cada cena e comparando esses *timestamps* com os *timestamps* resultantes de cada uma das técnicas segundo as metodologias já descritas. Portanto, as Tabelas 2, 3 e 4 apresentam os resultados de precisão e revocação em porcentagem (%), com a média aritmética dos telejornais de cada noticiário, por exemplo, JN representa a média das 5 edições analisadas e assim por diante. Ainda, no caso da técnica multimodal, foi criado um algoritmo para verificar se o *timestamp* $T_{(nome\_tcnica)}$ está no intervalo de um segundo e meio a mais ou a menos do *timestamp* da segmentação manual. Caso positivo, a variável `valor` (Equação 3) é acrescida segundo o valor de seu peso. Esse janela de tempo refere-se aos arredondamentos que cada técnica realiza ao retornar os *timestamps*.

Os resultados obtidos com as técnicas de extração de informação textual (Tabela 2, foram os melhores quando comparados somente com as outras técnicas aplicadas em separado[2]. 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 2: Resultados da técnica de texto

	P	R
JN	76.8	55.7
JH	69.7	30.6
JR	46.9	68.3
JG	79.2	52.0
$\bar{x}$	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.

Em relação as técnicas de áudio é possível observar, pela Tabela 3, que os melhores resultados de precisão foram obtidos com a ferramenta Audacity, com RMS conseguindo melhores resultados para revocação. Na prática, Audacity detém maior precisão na detecção de transição de cenas enquanto RMS detecta uma quantidade maior delas. Ainda, a detecção de silêncio foi utilizada em um estudo de caso como um indicador de presença de propagandas publicitárias. Como resultado, ambas as técnicas conseguiram atingir índices de mais de 65% de precisão, com destaque para RMS, com 72%.

Tabela 3: Resultados das técnicas sonoras

	Audacity		RMS	
	P	R	P	R
JN	54.4	40.0	33.6	45.6
JH	35.2	13.3	27.3	23.3
JR	34.3	26.1	19.2	36.2
JG	36.0	33.3	21.0	35.8
$\bar{x}$	40.0	28.2	25.3	35.3

Na Tabela 4 são apresentados os resultados das técnicas monomodais visuais e também da multimodal. Tais resultados consideram $m = 30$, ou seja, os 30 melhores resultados para as respectivas técnicas. Como observado desta tabela, o desempenho das técnicas de histograma global e

[2]O símbolo $\bar{x}$ representa a média aritmética simples de todos os telejornais.

wavelets foram similares e melhores que o histograma local para todos os grupos de telejornais, fato verificado também observando-se a média aritmética total dessas técnicas, com uma pequena vantagem para o histograma global.

Tabela 4: Resultados das técnicas visuais e multimodal

	Hist. Global		Hist. Local		Wavelets		Multimodal	
	P	R	P	R	P	R	P	R
JN	66.6	39.7	50.0	33.5	50.0	42.1	80.0	53.5
JH	58.6	46.2	54.0	43.0	61.3	48.1	70.0	55.2
JR	43.3	30.3	63.3	33.0	43.3	24.3	72.0	40.46
JG	58.0	52.24	53.3	41.3	58.6	53.0	74.6	67.14
$\bar{x}$	57.2	42.1	51.7	37.4	56.2	41.9	74.0	54.1

A principal característica do histograma global é considerar a intensidade das cores de toda a imagem, priorizando as cenas que possuem um único âncora justamente por possuir um plano de fundo estático e uniforme (Figura 4). Assim, é possível que a vantagem do histograma para as wavelets aconteça porque os telejornais, de modo geral, iniciam uma cena com a imagem de um único âncora e com plano de fundo estático. Outra vantagem desta técnica é a detecção de cenas que não apresentam o âncora, como as cenas de clima (última cena da Figura 4).

Figura 4: Imagens resultantes do histograma global para o Jornal Nacional de 09/03/2010

Os resultados das wavelets puderam identificar cenas de notícias onde dois âncoras estão presentes na imagem lado-a-lado ou cenas com a presença um único âncora com plano de fundo dinâmico (Figura 5). 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.

O histograma local não obteve a mesma eficácia, justamente por ter uma característica intermediária entre o histograma global e wavelets, pois como faz análise em blocos inclui características espaciais, no entanto, diferente das wavelets, tais características estão associadas à cor da imagem e não à textura. Mesmo não detectando mais de um âncora, esta técnica detectou âncoras com plano de fundo dinâmico, o que não acontece com histograma global.

Figura 5: Imagens resultantes das wavelets para o Jornal Nacional de 09/03/2010

Como observado na Tabela 4, a técnica multimodal desenvolvida consegue melhores resultados quando comparada com as técnicas monomodais em separado. Esta melhora ocorre porque o algoritmo de *ranking*, juntamente com o algoritmo de janela de tempo que leva em conta o arredondamento dos resultados das técnicas monomodais, consideram que os possíveis pontos fracos de uma técnica monomodal são compensados com os pontos fortes de outra técnica, considerando assim maior quantidade de segmentos semanticamente complexos e resultando em um desempenho melhor. Para efeito de comparação entre as técnicas de imagem e a técnica multimodal obtida com este trabalho, a Figura 6 apresenta seus desempenhos para a edição do Jornal da Globo de 04/03/2010. Observa-se que a multimodalidade (pontos na forma de círculos) consegue resultados melhores, fato que se comprova para os restantes do telejornais analisados.

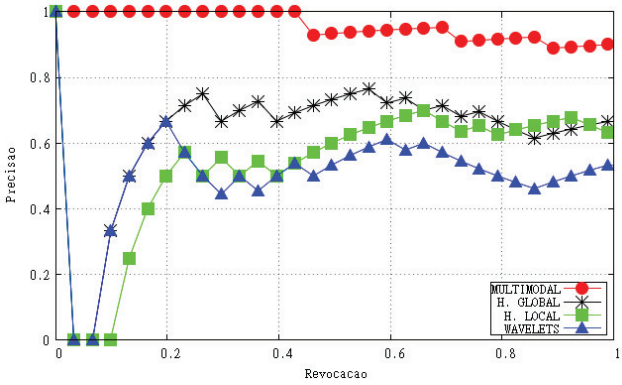

Figura 6: Comparação de desempenho entre as técnicas de imagem e multimodal

Por fim, a Figura 7 apresenta um exemplo de segmento semanticamente complexo de vídeo que possui transições de cenas detectadas pela técnica multimodal. Observa-se nesta figura três transições, a primeira detectada por uma técnica de histograma (local) e uma de silêncio (RMS), a segunda transição por silêncio (Audacity) e *closed-caption* e a última por histograma (global), silêncio (Audacity) e *closed-caption*. Em especial, a segunda transição representa a detecção entre segmentos semanticamente complexos sem a

imagem do âncora, pois a notícia relacionada às queimadas em florestas é seguida por outra notícia, mas relacionada a prisão de marginais, caracterizando uma mudança de assunto, ou seja, mudança de cena.

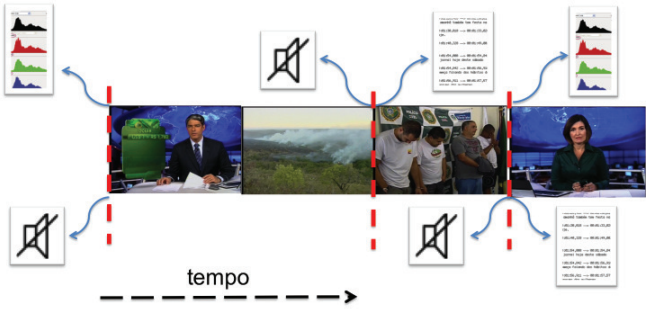

Figura 7: Exemplo de um segmento com transições de cenas capturadas pela técnica multimodal

7. CONCLUSÕES E TRABALHOS FUTUROS

Os resultados obtidos demonstram que a técnica desenvolvida tem como característica ser mais geral, pois considera segmentos semanticamente complexos nas notícias, além de levar em conta vinhetas, comerciais e cenas onde não aparecem âncoras, casos que não são tratados adequadamente em trabalhos relacionados. Como esperado, a integração de técnicas monomodais proporciona resultados melhores quando comparada com os resultados das mesmas técnicas aplicadas em separado, uma vez que as limitações de uma técnica são compensandas com vantagens de outras. Ainda, iniciou-se a criação de uma base de vídeos para testes de segmentação de cenas contendo riqueza de informação (contendo *closed-caption*) e telejornais completos.

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. Outra possibilidade é 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.

Agradecimentos

Os autores agradecem o apoio financeiro recebido das instituições FAPESP e CNPq.

8. REFERENCES

[1] J. M. Boggs and D. W. Petrie. *The Art of Watching Films*. Mayfield Publishing Company, 5th edition, 2000.

[2] L. Chaisorn, T.-S. Chua, and C.-H. Lee. A multi-modal approach to story segmentation for news video. *World Wide Web*, 6(2):187–208, June 2003.

[3] Y. Choi and J. Lee. Reliability and validity of scene unit coding in the visual content analysis. *Annual Meeting of the International Communication Association*, page 40, 2010.

[4] T.-S. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu. 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, pages 656–659, New York, NY, USA, 2004. ACM.

[5] F. Colace, P. Foggia, and G. Percannella. A probabilistic framework for tv-news stories detection and classification. In *Multimedia and Expo.IEEE International Conference on*, pages 1350 –1353, 2005.

[6] A. Hanjalic. *Content-Based Analysis of Digital Video*. Kluwer Academic Publishers, 2004. 193 pags.

[7] L. Hua-Yong and H. Tingting. Content-based story segmentation of news video by multimodal analysis. In *Fuzzy Systems and Knowledge Discovery, Sixth International Conference on*, pages 423 –426, 2009.

[8] W. Jianping, P. Tianqiang, and L. Bicheng. News video story segmentation based on naive bayes model. In *Natural Computation, 2009. Fifth International Conference on*, volume 6, pages 77 –81, 2009.

[9] Y. Li, W. Ming, and C.-C. Kuo. Semantic video content abstraction based on multiple cues. In *Proc. IEEE International Conference on Multimedia and Expo ICME 2001*, pages 623–626, 2001.

[10] H. Liu, T. He, and H. Zhang. Nbr: A content-based news video browsing and retrieval system. In *Technologies for E-Learning and Digital Entertainment*, pages 793–800. 2007.

[11] W. Liu, G. Yang, and X. Huang. Semantic features based news stories segmentation for news retrieval. In *Wavelet Analysis and Pattern Recognition, International Conference on*, pages 258 –265, 2009.

[12] Z. Liu, Y. Wang, and T. Chen. Audio feature extraction and analysis for scene segmentation and classification. *The Journal of VLSI Signal Processing*, 20:61–79, 1998.

[13] J. Magalhães and F. Pereira. Using mpeg standards for multimedia customization. *Signal Processing: Image Communication*, 19:437–456, 2004.

[14] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22:1349–1380, 2000.

[15] C. G. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25:5–35, 2005.

[16] J. Wang, L. Duan, Q. Liu, H. Lu, and J. Jin. A multimodal scheme for program segmentation and representation in broadcast video streams. *Multimedia, IEEE Transactions on*, 10(3):393–408, 2008.

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

[18] L. Zhao, S.-Q. Yang, and B. Feng. Video scene detection using slide windows method based on temporal constrain shot similarity. In *Proc. IEEE International Conference on Multimedia and Expo ICME 2001*, pages 1171–1174, 2001.

[19] M. Zhao, S.-Y. Neo, H.-K. Goh, and T.-S. Chua. Multi-faceted contextual model for person identification in news video. In *Multi-Media Modelling Conference Proceeding, 12th International*, page 8 pp., 2006.