UNIVERSIDADE FEDERAL FLUMINENSE

Combinando Imagem e Som para Detecção de Transições em Vídeos Digitais

por

Marcilene Scantamburlo Fonseca

Niterói
2006

Marcilene Scantamburlo Fonseca

Combinando Imagem e Som para Detecção de Transições em Vídeos Digitais

Dissertação apresentada ao curso de pós-graduação *strictu sensu* em Ciência da Computação, da Universidade Federal Fluminense – UFF, como pré-requisito para a obtenção do título de Mestre em Ciência da Computação, sob a orientação da professora Aura Conci e co-orientação do professor Paulo Sérgio Silva Rodrigues.

Combinando Imagem e Som para Detecção de Transições em Vídeos Digitais

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para obtenção do título de Mestre. Área de concentração: Computação Visual e Interfaces.

Avaliada em janeiro de 2006

Banca Examinadora

Profª Aura Conci – IC/UFF (orientadora)

Prof. Paulo Sérgio Silva Rodrigues – LNCC (co-orientador)

Profª Ana Cristina Bicharra Garcia – IC/UFF

Prof. Jauvane C. de Oliveira – LNCC

Prof. José Raphael Bokehi – IC/UFF

Agradecimentos

Primeiramente a Deus por ter me dado força para entrar na UFF e concluir a dissertação, um sonho que parecia tão distante. Também não posso esquecer da minha Mãezinha do Céu que esteve ao meu lado todo esse tempo.

A todos os meus amigos e familiares que sempre me apoiaram e que acreditaram em mim durante essa longa caminhada, em especial à Dona Zazá e maninho, por compreenderem minha ausência, ao amigo Jorge Valardan, que esteve ao meu lado durante todo o tempo, acreditando em mim e me dando apoio quando mais precisava e ao Hugo Theobald, pelos vídeos cedidos.

A amiga e orientadora Aura Conci que depositou em mim sua confiança e que nunca me deixou desanimar diante dos acontecimentos, sempre me dando forças.

Resumo

Um grande número de aplicações de vídeo digitais, como indexação, recuperação, análise, compressão, armazenamento e transmissão de dados, é realizado, primeiramente, baseado em detecção de transições de cena, às vezes, chamada de segmentação de vídeo. Por ser uma etapa base e estar longe de resolver todos os seus desafios, ainda é uma área em aberto, necessitando de muita investigação.

Embora a classificação de cenas pelo ser humano envolva diversos atributos, a utilização de cor, forma e textura ainda são os mais utilizados por modelos computacionais conhecidos. Por outro lado, pesquisas com processamento de sinais sonoros como voz têm ganhado bastante atenção, principalmente em sistemas multimídia. O processo de análise de cenas reais pode envolver uma ou mais características relacionadas a cada atributo, tanto visuais quanto sonoros.

O objetivo principal dessa dissertação é a análise dos resultados obtidos da detecção de transição de cenas, quando usadas características visuais combinadas com características de sinais sonoros associados às cenas. Essa detecção levou em consideração os aspectos de cor (baseado no cálculo do histograma de cores); e de forma (baseado no cálculo e histograma do mapa de bordas da imagem). Por sua vez, a detecção com base no conteúdo do som foi baseada na extração de 7 características, a saber: energia média de curta duração, taxa de cruzamento de zero, razão da energia de banda, magnitude do espectro delta, raiz quadrada média do somatório das intensidades dos sinais ao quadrado, razão dos valores de características altas e razão dos valores de características baixas.

Um modelo bayesiano foi utilizado para essa análise e os resultados mostraram que a inclusão do som à combinação de cor e forma pode melhorar o desempenho da detecção de transições.

Abstract

The large and growing volume of digital applications, such as video indexing, retrieving, analysis, compression, storing and data transmission, have being accomplished based on scene transition detection, sometimes called video segmentation. Since it is a basic step and it is far from a complete solutions, it is also a open area, needing much more for investigations.

Although scene classification by human concerns several attributes, the use of color, shape and texture are the more frequently used on computational models. On the other hand, researches on digital sound processing and voice signal have gotten much attention, mainly in multimedia systems. It is well known, however, that each attribution on processing of scene analysis can relies on several features at the same time, as visual as sonorous.

This dissertation proposes has as main goal the analysis of the results in automatic detection of scene transition using traditional visual features combined with sounding features to the scene. This detection regards the color aspects (based on color histogram), shape aspects (based on edge map histogram) and sound aspects, which is accomplished with 7 sound characteristics, namely: short time average energy, zero-crossing rate, energy band ratio, delta spectral magnitude, root mean square of square sum of signals, high sounds and low value characteristics ratios.

A Bayesian network model is used on this analysis and the results shows that the inclusion of sound in the model combining color and shape can improve the detection performance.

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

O avanço da multimídia e da *World Wide Web*, além das linhas de comunicação de alta velocidade e tecnologias para compressão de imagens, áudio e vídeo, tem incentivado o desenvolvimento de soluções que buscam armazenar, atualizar, manter, visualizar, recuperar e transmitir informações digitais.

Ao mesmo tempo, o volume de informações visuais e auditivas geradas vem aumentando a taxas elevadas, isto é, arquivos de áudios, imagens e vídeos disponibilizados na Internet, imagens de documentos corporativos, imagens e vídeos médicos dentre outros. Para armazenar e acessar estas informações de forma eficiente, técnicas de indexação baseadas em conteúdo estão sendo incorporadas aos atuais sistemas de informação. Neste caso, o conteúdo das imagens é modelado e representado usando conceitos de processamento de dados multimídia e ciência cognitiva.

A análise do conteúdo multimídia refere-se à compreensão de sua semântica[1]. Com todos esses avanços, ferramentas que possibilitem tais análises automatizadas estão se tornando indispensáveis. A semântica de um documento multimídia está contida em múltiplos aspectos que são, geralmente, complementares entre si.

A tarefa de segmentar, indexar e recuperar automaticamente dados audiovisuais, tem aplicações importantes em diversas áreas como na educação, no entretenimento, em segurança e, principalmente, no gerenciamento de arquivos multimídia. Por exemplo, há uma grande quantidade de materiais audiovisuais que devem ser arquivados em bancos de dados de filmes e televisão. Esses dados devem ser segmentados e

[1] Estudo das mudanças ou transições sofridas, no tempo e no espaço, pelo seu significado.

indexados apropriadamente, pois isso facilitará a recuperação de segmentos

de um vídeo para a edição de um documentário ou para a publicação de um clipe de vídeo. Um outro exemplo que pode ser citado é a coleção de vídeos de entretenimento de uma família, onde é conveniente para os usuários localizar e recuperar apenas segmentos de vídeo que lhes interessam.

A escolha das características a serem utilizadas é um dos pontos mais importantes e que tem demandado muito trabalho de pesquisa [LI01].

Essa metodologia apresenta duas grandes dificuldades, especialmente quanto ao tamanho das coleções de imagens. A primeira é a quantidade de trabalho despendido em uma anotação manual da imagem. A segunda é a essencial, pois resulta do conteúdo rico de uma imagem e a subjetividade da percepção humana, isto é, a mesma cena pode ser classificada de várias formas quando analisada por pessoas diferentes. A subjetividade da percepção e anotações imprecisas podem causar problemas no processo de recuperação.

A recuperação com base no conteúdo, por exemplo, assim substitui o método mais antigo de classificação, onde as características das imagens eram descritas textualmente, com a utilização de palavras-chave. Esse conteúdo textual é ainda armazenado e recuperado através de SGBD (Sistemas Gerenciadores de Bancos de Dados).

Nos anos 90, devido ao surgimento em grande escala das coleções de imagens, as dificuldades inerentes a anotação manual tornaram-se muito mais agudas. Para superar estes problemas foi proposta a recuperação de imagens com base no conteúdo. Isto é, além da anotação textual das imagens, estas seriam indexadas pelo seu conteúdo visual.

A indexação de imagens com base no conteúdo utiliza características internas, tais como cor e forma para classificação e recuperação. Inicialmente, as imagens passam por um pré-processamento; após, são extraídas suas características; em seguida, é feita a organização destes índices na forma de um vetor de características, que é finalmente usado quando da recuperação desta imagem em uma base de dados usando algum critério de decisão, (Figura 1.1.). A extração de características é o processo que tem gerado os

maiores desafios na área e também o que mais tem recebido atenção. Os problemas nesse processo vêm desde saber que características relevantes extrair até como obtê-las e armazená-las. Sabe-se, contudo, que são muitas as características que podem ser indexadas nesse tipo de sistema [RODRIGUES03, SMITH96].

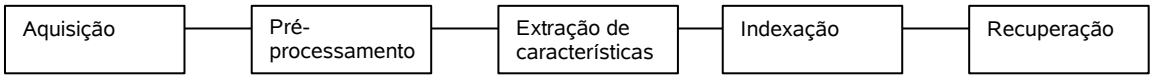


Figura 1.1: Etapas de recuperação de imagens com base no conteúdo

A recuperação de áudio com base no conteúdo também utiliza características internas, tais como volume, timbre, taxa de cruzamento de zero, centróide espectral, largura de banda, razão da energia de banda, magnitude do espectro delta, raiz quadrada média do somatório da intensidade do sinal ao quadrado, entre outras (Figura 1.2).

A segmentação e indexação automáticas baseadas no processamento computadorizado da análise do conteúdo multimídia é uma tendência evidente [TONG01].

O primeiro passo para a análise do conteúdo multimídia é a segmentação ou divisão de um documento. Isso significa selecionar um vídeo e dividi-lo (segmentá-lo) em cenas para que a mesma corresponda a uma unidade de história. As cenas são divididas em tomadas, lances ou *shots*[2], para que as características auditivas e/ou visuais permaneçam coerentes. Dependendo da aplicação, passos diferentes seguem o estágio da segmentação de áudio. Essa classificação semântica é a chave para gerar índices em formato de texto (Figura 1.2).

[2] Pequena parte de um vídeo ou filme, amostra de vídeo, lance ou tomada.

Finalmente, a sumarização é essencial em um sistema de recuperação de vídeo, que torna possível ao usuário procurar rapidamente dentro de um

grande conjunto de itens a resposta para sua procura. Além do resumo de um texto com o conteúdo do vídeo, alguns recursos de áudio e vídeo dão ao usuário uma idéia melhor dos personagens, lugares e estilo do vídeo.

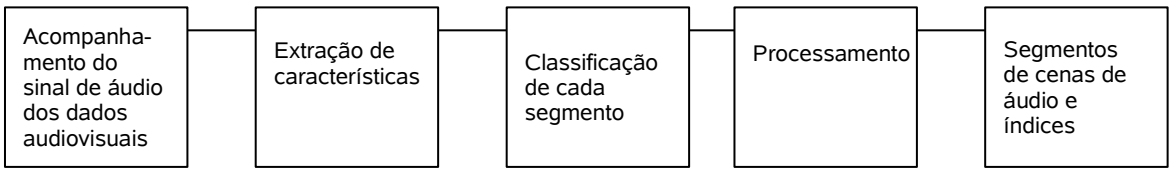


Figura 1.2: Indexação e segmentação automática de dados audiovisuais baseada na análise do conteúdo de áudio

Os passos acima não são exclusivos, mas podem partilhar com alguns elementos básicos ou ainda serem interdependentes, por exemplo: sumarização e indexação podem exigir a extração de alguns quadros ou *frames*[3] chave dentro de cada cena ou lance que melhor reflita seu conteúdo visual. Da mesma forma, a classificação e a segmentação da cena são dependentes umas das outras, porque os critérios de segmentação são determinados pelas classes de definição das mesmas.

O ponto crucial para o sucesso desses passos é a extração de características auditivas e visuais apropriadas. Tais características não são somente úteis como índices de baixo nível - que seria a recuperação através do conteúdo textual - mas também oferecem uma base para a comparação entre cenas e tomadas. Tal comparação é necessária para segmentação e classificação das cenas e lances como também para a escolha de amostras de quadros/clipes para a sumarização [CONCI02].

---

[3] Quadros. A taxa de quadros ou *frames* indica o número de *quadros por* segundo. 24 *frames*/s para filme, 25 *frames*/s para vídeo europeu, 29,97 *frames*/s para NTSC vídeo à cores, 30 *frames* para NTSC P&B e para a maioria das aplicações musicais.

Para a recuperação de quadros ou cenas em um banco de dados, uma imagem/cena exemplo pode ser usada a fim de serem recuperadas quadros similares segundo o vetor de características. Também, imagens e cenas podem passar por um processo de análise a fim de serem agrupadas segundo algum critério de similaridade, sem a utilização de uma imagem/cena busca [CONCI02]. O mesmo pode ocorrer com um clipe de vídeo, onde o usuário possa recuperar um clipe que seja visual ou auditivamente semelhante com um clipe exemplo.

1.1. Objetivo da dissertação

Este trabalho é o segundo na linha de pesquisa baseada na recuperação de imagens por conteúdo, do IC/UFF. O trabalho anterior [CASTRO98] dedica-se à busca com base no conteúdo cromático e dedicou-se a busca do melhor espaço de cor para melhor caracterização do histograma e da métrica usada [CONCI99].

O objetivo da presente dissertação é combinar os métodos de Recuperação de Imagens com Base no Conteúdo - RIBC (cor e forma) com os de recuperação de áudio com base no conteúdo, no sentido de agrupar quadros e tomadas que façam parte da mesma cena de um vídeo, indicando pontos de transição de uma cena para outra (indexação de cenas). Em relação aos trabalhos anteriores da linha, houve um crescimento de imagens para vídeo e também do número de características estudadas.

Uma vez que as características de som e imagem são altamente correlacionadas, é interessante combiná-las através de um modelo probabilístico. Assim, este trabalho usa o modelo proposto em [RIBEIRO00] de redes bayesianas para indexação de textos, e complementado por [RODRIGUES05], para indexação de imagens baseada em cor, forma, textura e relações espaciais. Neste trabalho, propõe-se a inclusão de características de som no modelo bayesiano utilizado. Para mostrar sua eficiência, foram realizados experimentos em cinco vídeos digitais compostos por mais de

40.000 quadros com cerca de 60 transições de cenas. Resultados mostram que a maioria das transições é detectada e o número de falsos positivos é pequeno, entretanto há um número de falsos negativos que deve ser levado em consideração.

1.2. Organização do Trabalho

No Capítulo 2, são apresentadas as contribuições anteriores, os trabalhos relacionados com a área de recuperação, com base no conteúdo de imagens e sons e indexação de vídeo.

No capítulo seguinte, é mostrada a teoria necessária para o entendimento sobre cada técnica utilizada desde as características de cor, forma e som, até o modelo vetorial, a medida de similaridade e a teoria bayesiana, base da presente dissertação.

Logo após, no Capítulo 4, é apresentada a metodologia utilizada e como serão extraídas as características selecionadas - cor, forma e som. Nesse mesmo capítulo, é explicado o relacionamento entre as características selecionadas.

No Capítulo 5, são apresentados os experimentos; os resultados da implementação e as novas contribuições e, finalmente, no Capítulo 6, são expostas as conclusões como considerações finais e as sugestões para possíveis trabalhos futuros.

2. Trabalhos Relacionados

2.1 Recuperação de Imagens

As pesquisas em Recuperação de Imagens com Base no Conteúdo (RIBC) têm uma história relativamente recente; a maioria dos trabalhos data de meados da década de 90 em diante [HERMES95, SCHEIRER97, CONCI99, GOUYON00, LI01, LU02, EAKINS02, TZANETAKIS02].

Desde então, é uma área de interesse crescente, que envolve técnicas adaptadas de outras áreas da Ciência da Computação como Reconhecimento de Padrões e Inteligência Artificial.

Dada essa interseção de soluções, nem sempre é possível distinguir quando o objetivo é análise ou recuperação de imagens com base no conteúdo. Entretanto, de acordo com as diversas técnicas utilizadas, [EAKINS02] apresenta uma classificação interessante da área de RIBC:

1. classificação automática de cenas, que tipicamente usam métodos estatísticos;

2. classificação automática de objetos, usando uma das seguintes abordagens:

2.1 técnicas baseadas em modelos de objetos previamente armazenados;

2.2 técnicas estatísticas, semelhantes àquelas usadas para classificação de cenas;

3. técnicas de realimentação de informações através de usuários.

Um dos primeiros sistemas para RIBC foi o Sistema IRIS [HERMES95], que usa uma combinação de cor, textura e relação espacial entre as regiões para interpretar uma imagem, gerando descrições do tipo montanha, floresta, lago etc., que servem de entrada para um sistema com interface baseada em

texto (Figura 2.1).

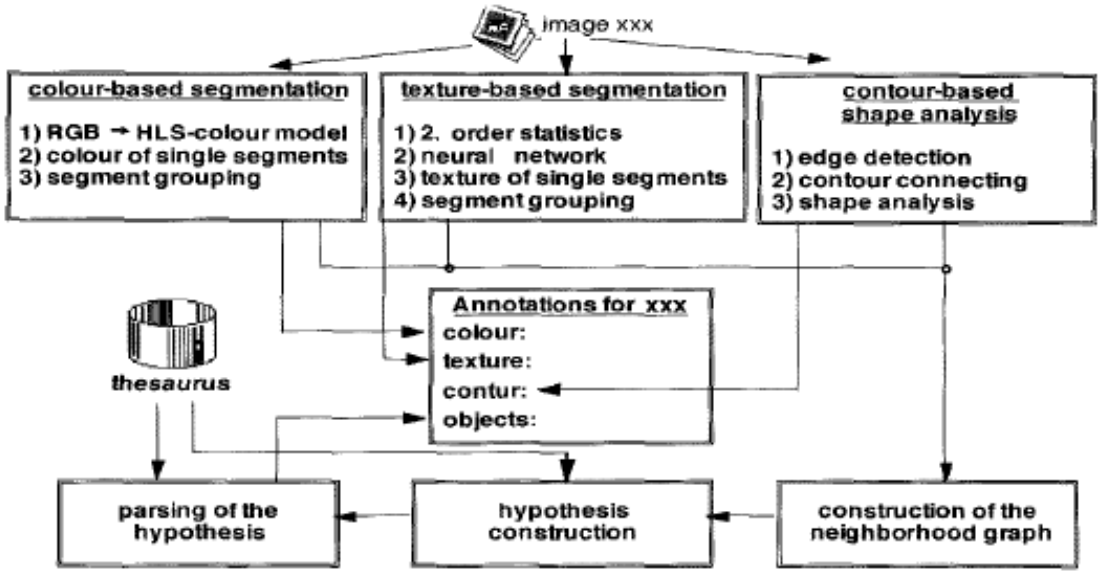


Figura 2.1: Arquitetura do sistema IRIS [HERMES95]

O sistema apresentado por [OLIVA99] usa filtros de Gabor[1] sobre atributos de bordas como característica global da imagem para separar as imagens em duas classes: artificial e natural e ainda ambiente fechado e aberto. Em [SZUMMER98] é usada uma combinação entre histogramas de cor utilizando o espaço HSV (*Hue, Saturation, Value)*, textura e Transformada Discreta de Cossenos para treinar um classificador baseado na regra de decisão do vizinho mais próximo para distinguir entre imagens ao ar livre e em ambientes fechados.

Em [LIPSON97], é proposta uma abordagem diferente baseada em uma análise semântica qualitativa, usando reconhecimento de padrões e uma combinação de distribuição de cor para classificar imagens como montanhas ou campos.

---

[1] Segundo [LEE96], o conjunto de funções que compõem os filtros de Gabor modela os campos receptivos das células do córtex visual humano, localizado no cérebro.

[VAILAYA98] apresenta um modelo de classificador bayesiano para agrupar imagens em um número de categorias, incluindo cidades e paisagens e florestas e montanhas, usando vetores gerados por quantização vetorial a partir de uma análise de momentos, gerando coeficientes de cores e filtros de Gabor.

Em [MATHIAS98] a influência do nível de discretização no uso do histograma e o espaço de cor usada é considerado. O uso desse histograma como vetor de características em um sistema baseado na recuperação de imagens semelhante à uma imagem exemplo, quando nesta recuperação é baseada o conteúdo cromático é tratada em [CONCI99]. Como diversas métricas podem influenciar no critério de decisão sobre qual a melhor recuperação é considerada em [CONCI00] e [CONCI01]. A combinação dos diferentes critérios em uma metodologia de indexação baseada em cor relacionada a uma base de dados única com cinco diversos critérios de decisão são estudados em [CONCI02]

No trabalho de [ZHOU01], é mostrado um algoritmo que extrai características exclusivamente do mapa de bordas da imagem original. O algoritmo percorre o mapa e, durante a varredura, extrai características espaciais consideradas globais da imagem. Esses resultados mostram-se eficientes com relação à precisão.

De uma forma geral, podemos dizer que os problemas dominantes na área de RIBC são: a extração de características, a indexação, a definição subjetiva do que é relevante ou não como resposta a uma requisição do usuário e a combinação de características heterogêneas que possam produzir os melhores resultados. Considerando, então, a natureza estatística da maioria dos problemas citados, o uso de modelos estatísticos (como modelos bayesianos) parece ser um caminho natural para a sua solução.

Em 1991, o trabalho de [TURTLE91], batizado de *Inference Network Model*, foi o primeiro a utilizar redes bayesianas para recuperação de informação. Seu objetivo, no entanto, é exclusivamente recuperação de textos. Pouco depois, um modelo mais genérico, chamado *Belief Network Model for*

Information Retrieval, foi proposto por [RIBEIRO96]. Esta variante introduz evidências de consultas passadas em uma rede bayesiana com o objetivo de melhorar a qualidade das respostas. Com esse mesmo objetivo, procurando maximizar o número de documentos relevantes e minimizar o número de documentos não relevantes na resposta a uma consulta, [SILVA00] propôs um modelo bayesiano capaz de representar em um mesmo arcabouço os três modelos clássicos de recuperação de informação até então vigentes [YATES99] (booleano, vetorial e probabilístico), usando ciclos de realimentação de relevantes e alternativas de similaridade consulta-consulta.

Em [COELHO04] é apresentada uma técnica para recuperação de imagens usando uma rede bayesiana inspirada nos mesmos modelos para busca em texto que já havia sido testado com sucesso em [RIBEIRO96, RIBEIRO98, RIBEIRO00, SILVA00]. No entanto, o conteúdo de informação utilizado ainda era um conjunto de evidências de texto (não pertencentes, portanto, ao conteúdo semântico da imagem) como *links* e *meta tags* de páginas HTML (*HyperText Markup Language*) de onde as imagens eram retiradas.

Recentemente, em [RODRIGUES03] é apresentado um novo algoritmo, chamado GRAS, acrônimo para *Graph Region Arrow Shot*, para extração de características do relacionamento espacial entre regiões da imagem (objetos e cenas). Essas características são adicionadas ao espaço vetorial no intuito de agregar as informações globais às primitivas locais.

Em [RODRIGUES05], o modelo proposto por [RIBEIRO00] para recuperação de textos e por [COELHO04] para recuperação de imagens baseada em texto é estendido para recuperação de imagens com base no conteúdo de cor e forma.

2.2 Recuperação de Áudio

Os avanços no reconhecimento automático da fala (ASR – *Automatic Speech Recognition*) estão aumentando o interesse na classificação geral de

dados de áudio (GAD – *General Áudio Data*). Muitos esquemas de classificação de áudio têm sido investigados nos últimos anos, eles geralmente se diferem em duas maneiras: na escolha da classificação e na série de características acústicas usadas [FOOTE99].

Uma série de características acústicas usadas nos sistemas atuais incluem classificadores com base no modelo Gaussiano, redes neurais, árvore de decisão e o modelo oculto de Markov [JAVED00].

[SCHEIRER97] avalia várias combinações de 13 características temporais e espectrais usando diversas estratégias de classificação. Consegue precisão de 90% considerando fala e música, mas apenas 65% quando consideraram separadamente fala e música e fala e música simultâneamente.

[SPINA96] usa 14 MFCC (*Mel-Frequency Cepstral Coefficients*) para classificar o áudio em 7 categorias: fala em estúdio; fala em campo; fala com fundo musical; fala com barulho; música; silêncio e barulho, cobrindo o restante dos outros padrões.

Em muitos casos, a seleção das características é mais crítica que a performance da classificação. [LI01] avalia um total de 143 características de classificação para o problema da classificação contínua (GAD) e chega a 7 categorias: silêncio, fala de uma pessoa, música, ambiente barulhento, várias pessoas falando, fala e música simultâneamente e fala com barulho.

[ZHANG00] propõe um esquema de segmentação de áudio baseado na divisão em duas partes: a primeira consiste em discriminar o que é ou não voz. Para isso ele utiliza o algoritmo de classificação KNN *K-Nearest Neighbor,* vizinho mais próximo de K, baseado na taxa de cruzamento de zero.

[ZHU03] classifica os segmentos de áudio em quatro tipos de dados semânticos: silêncio, som ambiental, voz e música. Ele se utiliza da taxa de energia baixa da freqüência (LER) para descrever o número de quadros com energia mais baixa do que determinado nível baseado no BIC (*Bayesian Information Criterion)* com o objetivo de detectar mudanças do microfone no

áudio.

O método apresentado por [PARASKEVAS03] faz uso efetivo do espectograma, da distribuição de *Wigner-Ville* e das matrizes de co-ocorrência. As imagens são comprimidas utilizando métodos estatísticos antes de serem combinadas numa única matriz de características para serem então, inseridas no algoritmo classificador.

[TONG01] propõe um esquema para a automatização da segmentação e classificação de dados audiovisuais baseados na análise do conteúdo de áudio. Quatro tipos de características são extraídas, função-energia de curto tempo, média da taxa de cruzamento de curto tempo, freqüência fundamental de curto tempo e espectro de trilhas de pico. Eles realizam uma análise morfológica e estatística das curvas temporais dessas características para revelar as diferenças entre diferentes tipos de áudio. Um procedimento heurístico baseado em regras é construído para segmentar e classificar os sinais de áudio com essas características.

2.3 Transições de Vídeo

Vários grupos de pesquisa utilizam a informação de áudio e vídeo para detectar as fronteiras de uma cena [YOSHITAKA01, SUNDARAM00, MURAMOTO00, JIANG00]. Existem alguns métodos de busca denominados *query-by-content-object* onde objetos são procurados em um vídeo, analisando as relações temporais e espaciais entre tais objetos, entretanto métodos como esse são computacionalmente custosos. A análise de seqüências de imagens do vídeo são as mais eficientes. Nesse caso, o agrupamento de cena é feito medindo a similaridade entre os quadros.

Uma grande variedade de algoritmos para detecção de transição de cenas tem sido pesquisada, as transições abruptas são as mais estudadas e é uma das transições mais comuns em um vídeo. A maioria dos métodos se baseia em apenas uma característica de vídeo. [NAGASAKA92] e [ZHANG93]

propõem o uso de esquemas de iluminação, entretanto a informação espacial das imagens é descartada nos histogramas.

[FISHER95] usa estatística de amplitude de freqüência para classificar vídeo clipes em notícias, esportes e comerciais, já [LIU98] segmenta e categoriza vídeos em notícia, basquete, futebol, comercial e segmentos diversos combinando características visuais e auditivas de tal modo que a segmentação final está baseada na junção dos limites dos resultados visuais e sonoros.

Após testar diversas características [ZHU03] classifica o áudio em 4 segmentos: fala, música, som ambiente e silêncio. Devido à facilidade de segmentação e implementação, opta pela utilização das seguintes características: energia média de tempo curto, taxa de cruzamento de zero, freqüência fundamental e freqüência do centróide. Utiliza um algoritmo que percorre a estrutura do vídeo, selecionando tomadas baseados nos termos DC (*Dublin Core)*. Os termos DC dos quadros I estão diretamente disponíveis nas trilhas MPEG (*Moving Picture Experts Group)*. A técnica de reconstrução é custosa computacionalmente, entretanto [SHEN96] propôs um algoritmo rápido onde os termos DC dos quadros B e P são reconstruídos eficientemente. Imagens DC, que são os valores médios do bloco *(i, j)* da imagem, podem ser construídas a partir dos termos DC.

[KORPI03] melhora a qualidade do vídeo para detectar mudança de cenas. A melhora está baseada na cor calculada através do histograma RGB. A meta é agrupar as tomadas sucessivas, no sentido de formar cenas semanticamente corretas baseadas nas dicas visuais e auditivas, usando regras de interação.

[ZHANG99] desenvolve um sistema de classificação de áudio baseado em regras heurísticas e na aplicação da segmentação de vídeo. A detecção das transições de cena depende de pesos atribuídos a cada característica. Já [ZABIH99] propõe um método de detecção de cena baseado no contorno, entretanto nenhuma informação foi dada sobre a confiabilidade dos valores de similaridade calculados quando o número de pontos do contorno era muito

baixo.

[THOMPSON00] introduz um método que aplica modelagem de estatística linear para estimar mudança de cena levando em conta as mudanças na visão da imagem (tomada da câmera). Nesse método é descrita a variação quantitativa de um pixel em torno de seus vizinhos (fronteira) espacial e temporalmente, com um modelo de probabilidade linear. Depois são estimados os modelos de parâmetros com mínimos quadrados, é calculado o diagnóstico de modelagem para cada fronteira e o resultado é mostrado como imagens, de modo a avaliar a mudança de cena e o efeito de visão ou da tomada de câmera.

[REN02] utiliza uma rede neural e um classificador *k-nearest* (vizinho mais próximo) para modelar transições em vídeo baseado em um número de medidas estatísticas derivadas das imagens do vídeo. As transições que podem ser modeladas são as graduais, e incluem efeitos como: *cut, fade in, fade out, dissolve, tilt up, tilt down, pan left* e *pan right*.

[VASCONCELOS97] introduz modelos estatísticos para dois componentes importantes para extrair características que permitam classificação semântica: duração da tomada e atividade; e demonstra a utilidade desses modelos introduzindo uma formulação bayesiana para o problema da segmentação de tomadas da câmera.

Em [JAVED00] é mostrado um método para remover comerciais de vídeos de entrevistas e para segmentar entrevistas em tomadas separadas. O artigo concentra-se na informação contida nas transições de cena mais do que nos frames individualmente.

[RASHEED02] utiliza uma técnica para reunir as tomadas em cenas, transformando o problema de particionamento em grafos. Isso é feito construindo um grafo não direcionado, chamado de *shot similarity graph* (SSG), de modo que cada nodo de grafo representa uma tomada e suas ligações possuem pesos.

No próximo capítulo, as técnicas utilizadas para a detecção de transição de cena, considerando os aspectos de cor e forma dos quadros. As informações auditivas serão apresentadas e detalhadas também no Capítulo 3, que se segue.

3. Técnicas Utilizadas

3.1 Segmentação de Vídeo

A segmentação de vídeo em cenas ou tomadas (*shots*) é um passo fundamental para analisar o conteúdo de uma seqüência de vídeo e para o acesso, recuperação e procura eficientes em grandes bancos de dados de vídeos. Uma cena é composta por uma ou mais tomadas sucessivas, organizadas por certas regras semânticas; uma tomada é composta por uma seqüência de quadros consecutivos do início ao final da gravação de uma posição de câmera. Desta forma as cenas podem ser mais semânticas que as tomadas e o conteúdo das diversas imagens interiores a uma tomada é similar (Figura 3.1) [LIENHART97]

O processo de segmentação de vídeo em tomadas de câmera é caracterizado pela busca das fronteiras entre duas tomadas consecutivas, analisando a similaridade entre os quadros. Os diversos tipos de algoritmos para detecção dessas fronteiras possuem diferentes abordagens para caracterizarem tal similaridade.

O áudio contribui com muitas informações complementares que melhoram a detecção da transição de cenas como, apenas ouvindo um jornal, é possível definir uma transição entre uma notícia e outra ou entre uma notícia e um comercial ou mesmo dois comerciais, da mesma forma que é possível, apenas ouvindo um jogo de futebol, identificar mudanças de cenas.

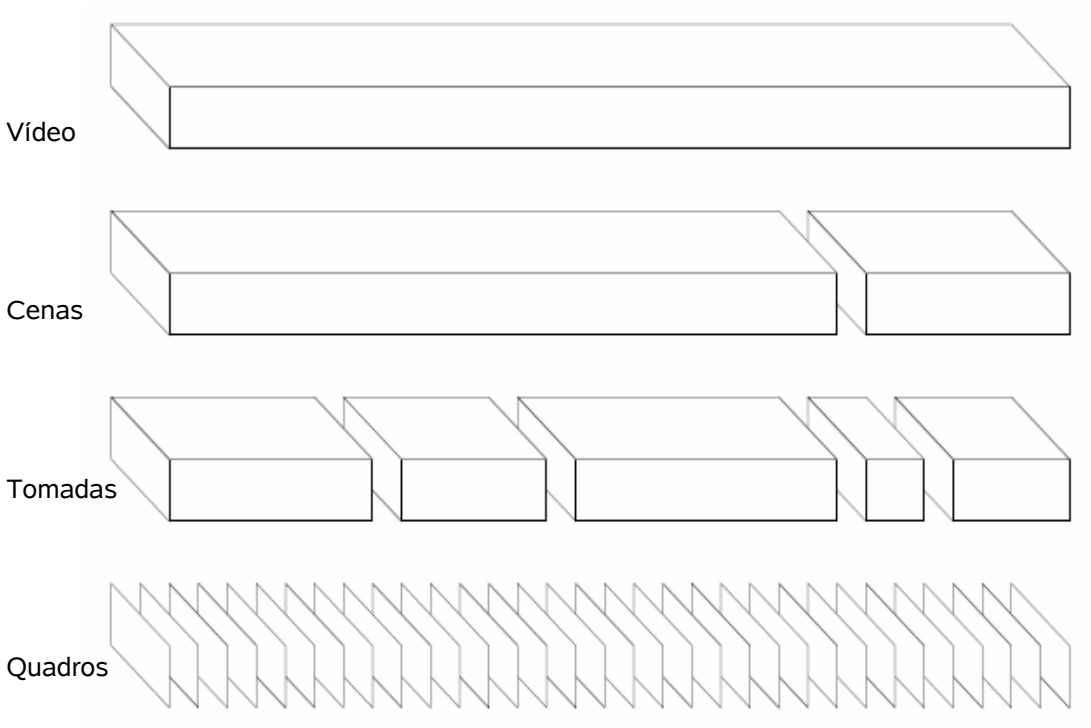


Figura 3.1: Estruturação de vídeo, como visto em [LIENHART97]

O objetivo final da segmentação de vídeo é agrupar automaticamente tomadas de modo que o ser humano identifique como sendo da mesma cena. Alguns trabalhos também definem cenas como sendo unidades de uma história, segmentos de história ou parágrafos de vídeo. Ao se usar a terminologia de um filme, uma tomada, também conhecida como lance, é uma imagem com um único *frame* móvel ou estático, a cena é geralmente composta por um pequeno número de lances inter-relacionados. Uma tomada pode ser um grupo de quadros que contém características visuais consistentes, incluindo cor, texto e animação. Tipicamente, a direção da câmera e o ângulo da vista definem um lance. Quando a câmera foca para a mesma cena por ângulos diferentes ou para regiões diferentes de uma cena pelo mesmo ângulo, é possível ver diferentes tomadas da câmera.

Uma vez que os lances são caracterizados pela cadência de algumas características visuais de baixo nível, não é muito difícil separar um vídeo em tomadas. Por outro lado, o agrupamento de uma tomada em cenas, depende do julgamento subjetivo da correlação semântica. Tal agrupamento requer a compreensão do conteúdo semântico do vídeo através da análise do agrupamento de características de áudio e vídeo. No entanto, é possível reconhecer lances que estão relacionados a locais ou eventos sem efetivamente utilizar análises de alto nível de questões semânticas. Da mesma forma que a indexação de imagens, trabalhos mais antigos em segmentação de vídeo focavam o uso de informação visual, entretanto, reconhecendo a importância do áudio nos segmentos de vídeo, esta dissertação utiliza segmentação de conteúdo usando a análise de imagens (cor e forma) e as características de áudio.

3.1.1 Tipos de transição

As transições mais comuns e simples entre duas tomadas de câmera são as transições abruptas, também conhecidas como corte seco, que representam, na verdade, a transição sem nenhum efeito de edição. Nesse tipo de transição, o último quadro da primeira tomada é seguido pelo primeiro quadro da segunda tomada. Existem também as transições graduais que resultam de combinações cromáticas ou espaciais de quadros das duas tomadas, sendo as principais: *dissolve, wipe, fade in* e *fade out*.

3.2 Processamento das características de cor

A escolha de cor e forma foi motivada pela sua larga utilização na literatura e simplicidade de implementação. São muitas as propostas na literatura que apresentam uma combinação de características de cor e forma,

destacada na Seção 3.3, para sistemas de indexação, clusterização[1] e RIBC. Uma vez que essas características estão estritamente relacionadas com o significado dos dados envolvidos na literatura de recuperação de informação, elas também são chamadas de evidências semânticas, ou simplesmente evidências [RODRIGUES03].

A grande maioria dos trabalhos de recuperação de imagens com base no conteúdo utiliza o processamento da cor, normalmente feito através do cálculo do histograma de cores quantizadas [MATHIAS98].

Neste trabalho, a primeira etapa do processamento da característica de cor, é dada da seguinte forma: cada *frame* I_j é convertido do sistema RGB para o HSV [CONCI99]. Esse sistema reflete melhor a percepção de cor pelo Sistema Visual Humano se comparado com o sistema RGB [CASTRO98].

Depois é feita a quantização do espaço de cor da imagem com o objetivo de diminuir o espaço em disco requerido e a quantidade de processamento necessária. No entanto, é sabido que quantizar pode levar a perdas de informações relevantes e detalhes importantes [CONCI99].

---

[1] Clusterização é um processo de agrupamento de dados que apresentam um elenco de propriedades similares, agrupadas de forma distintas daquelas presentes em outros dados não pertencentes ao mesmo agrupamento.

Foram adotados 18 tons para representarem a matiz (*Hue*), de acordo com as reproduções de diferentes cores possíveis; 3 valores cuja saturação (*Saturation*), demonstra a quantidade de branco presente na cor; e 3 valores para a intensidade (*Value*), representando a quantidade de luz presente [YOO02]. Com esses valores, é possível codificar 18 x 3 x 3 = 162 tipos de cores para cada *pixel*. Cada uma dessas 162 combinações é um elemento de cor quantizado no histograma da imagem. As faixas de valores possíveis adotados para H, S e V são limitados por: $H = 1, 2, \ldots, 18$; $S = 0, 0.5, 1.0$ e $V = 0, 0.5, 1.0$.

Assim a representação de uma imagem I_j no sistema HSV é feita por um histograma de 162 entradas possíveis, chamado de "histograma-162". Para haver invariância com relação ao tamanho da imagem este histograma será dividido pelo número de *pixels* da imagem, o que é chamado de normalização pelo tamanho da imagem e descrito pela Equação 3.1.

$$f_{i,j} = \frac{H_j(i)}{Lin \times Col} \tag{3.1}$$

Nessa equação, $f_{i,j}$ é o valor da imagem i no histograma de cor H_j e representa a freqüência do elemento i no histograma-162 dessa mesma imagem, Lin e Col representam o número de *pixels* por linha e coluna da imagem, respectivamente.

Os termos mais freqüentes pouco ajudam na classificação, uma vez que tendem a representar características comuns a todos os quadros do vídeo. O mesmo pode ser verdade para os termos menos freqüentes, pois, uma vez que são pouco comuns, tendem a maximizar a separação interclasses e minimizar a separação intraclasses. Uma forma de expressar quantitativamente estas relações é através da freqüência inversa, dada pela Equação 3.2, onde N é o número de quadros do vídeo e n_i é o número de imagens em que o elemento i aparece.

$$f_{inv} = \log \frac{N}{n_i} \tag{3.2}$$

Pode-se observar que f_{inv} é decrescente com relação a n_i, expressando portanto o fato de que os termos menos freqüentes são mais significativos para a classificação. Assim, a cada cor k_i no frame j pode ser associado um peso, como na Equação 3.3, o qual será usado para ponderar a importância da cor k_i para a recuperação.

$$w_{i,j} = f_{i,j} \times ¿ \log \frac{N}{n_i}$$

(3.3)

A indexação de imagens com base no conteúdo pode ser interpretada como um problema de clusterização de dados, como pode ser visto na Figura 3.2.

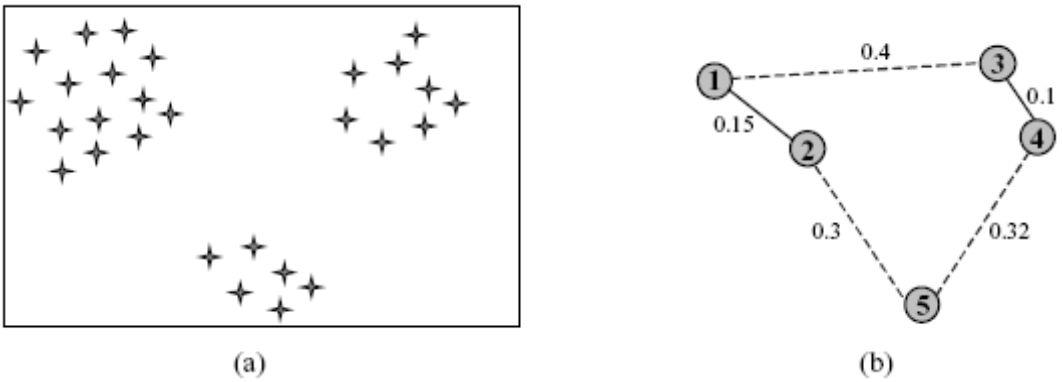
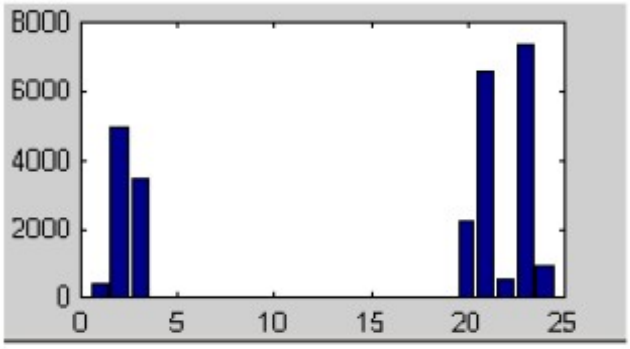


Figura 3.2: (a) Conjunto de quadros a serem reconhecidos e clusterizados, contendo três regiões de alta concentração.
(b) imagens rotuladas e os respectivos pesos das conexões.

Os histogramas são utilizados porque, além de serem fáceis de computar, independem da posição ou tamanho da imagem. Entretanto, quadros diferentes podem possuir histogramas de cores iguais e quadros semanticamente iguais podem possuir histogramas de cores diferentes. Um exemplo pode ser visto na Figura 3.3 onde analisando visualmente, a imagem (a) é mais semelhante a (c) do que a (e), entretanto, dependendo da métrica usada pode ocorrer que os histogramas, a imagem (a) sejam quantitativamente mais semelhante a (e) do que a (c), devido ao histograma (b) poder ser numericamente mais semelhante ao histograma (f) do que ao (d), em uma

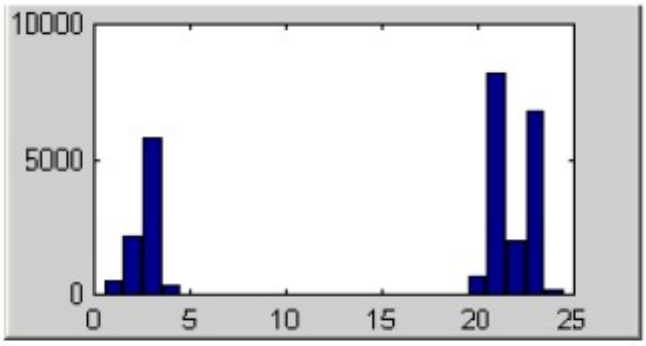
dada métrica.

(a)



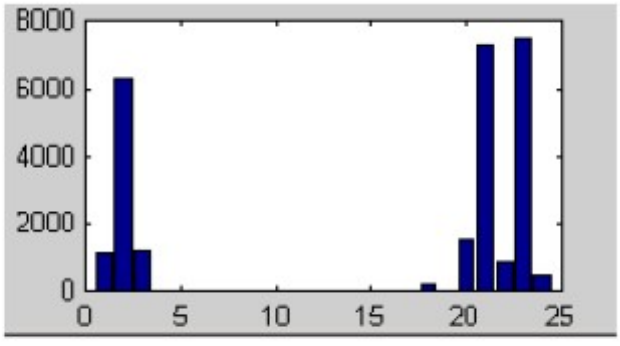
(b)

(c)



(d)

(e)



(f)

Figura 3.3: Exemplo de histogramas de imagens quantizadas para 24 cores no espaço HSV

3.3 Processamento das características de forma

A escolha da característica de forma, assim como a cor, foi motivada pela sua larga utilização na literatura e simplicidade de implementação. Essa característica diz respeito às formas (*shapes*) dos objetos que compõem uma imagem. A maneira mais simples de identificar tais objetos é através da binarização, em que os *pixels* com valores abaixo de um determinado limiar são transformados em 0 (preto) e os *pixels* acima desse limiar são transformados em 1 (branco)[2]. A Figura 3.4 ilustra um exemplo de binarização com limiar igual a 128.

[2] É importante observar que a binarização não funciona bem em diversas situações como em imagens com baixo contraste entre as regiões a serem separadas, ou com iluminação não uniforme.

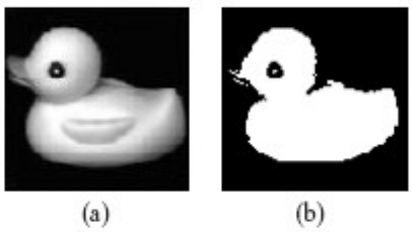


Figura 3.4: Exemplo de binarização. (a) Imagem em tons de cinza. (b) Imagem binária

Os filtros detectores de bordas - como *Canny*, *Prewitt*, *Sobel*, entre outros, também permitem uma indicação dos limites de um objeto dentro de

uma imagem [PRATT91]. Para esses detectores, uma borda é caracterizada por uma mudança brusca de um nível de cinza para outro, representando a fronteira entre duas regiões. A Figura 3.5 ilustra um exemplo de detecção de bordas, no qual é obtida apenas a informação do contorno do objeto.

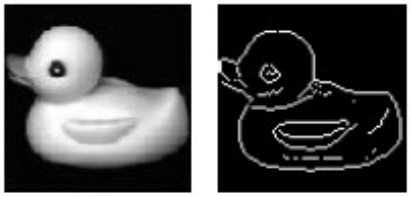


Figura 3.5: Exemplo de extração de bordas. (a) Imagem em tons de cinza.

(b) Imagem de bordas

Uma outra forma de segmentar imagens em objetos componentes é a partir de técnicas baseadas em regiões. A técnica conhecida como Crescimento de Regiões [BRICE70], por exemplo, agrupa iterativamente *pixels* com propriedades similares, como, por exemplo, nível de cinza, textura, cor dentre outros, em regiões maiores, até alcançar um determinado critério de parada.

Existem RIBCs - como o SAFE [SMITH97], por exemplo – onde é possível realizar consultas baseadas em regiões (ou objetos) independentes da imagem. Esse tipo de segmentação, que independe da localização física do objeto na imagem, é chamada de *espacial*.

Na extração da característica de forma deste trabalho, primeiramente o quadro é convertido para tons de cinza, logo após é aplicado a filtragem de Sobel [PRATT91] e finalmente é calculado o mapa de bordas.

3.3.1 Filtro Sobel

Os filtros são modos de se manipular as matrizes de *pixels* de modo a obterem-se informações desejadas ou eliminar aquelas que prejudicariam o correto funcionamento da aplicação.

A maioria dos filtros no domínio espacial da imagem se baseiam em uma matriz de convolução. Essa matriz é responsável por definir, para cada elemento, como este se relaciona com os seus vizinhos. O valor de cada *pixel* da imagem obtida após a filtragem será determinado, então, como resultado das operações definidas pela matriz de convolução para os elementos correspondentes da matriz original.

O filtro de Sobel é um filtro não-linear para realçar bordas e representa uma aproximação à função de Sobel, além de ser o mais conhecido detector de bordas. Envolve a aplicação de dois *kernels* direcionais e a combinação dos resultados (Equação 3.4). O filtro Sobel de detecção de contornos verticais (3x3) é dado pela máscara da tabela 3.1 e o filtro de detecção de contornos horizontais (3x3) é dado pela máscara da tabela 3.2:

Tabela 3.1: Filtro de Sobel (3x3) vertical

-1	0	1
-2	0	2
-1	0	1

Tabela 3.2: Filtro de Sobel (3x3) horizontal

-1	-2	-1
0	0	0
1	2	1

$$Sobel = \sqrt{(Dx)^2 + (Dy)^2} \tag{3.4}$$

Primeiramente os quadros em TIF são passados para uma resolução menor (180 x 120); logo após são passados para tons de cinza; o filtro de Sobel é aplicado na horizontal e na vertical; o cálculo das tangentes é realizado e, por último, é gerado o histograma (Figura 3.6).

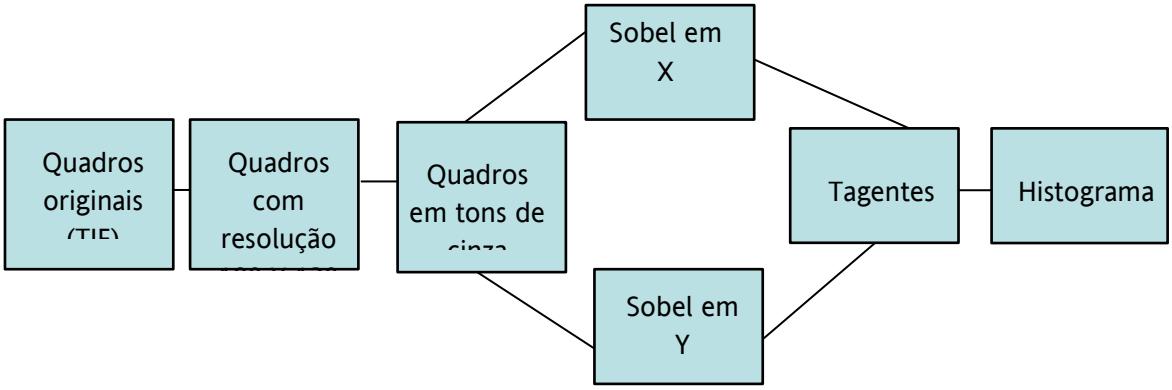


Figura 3.6: Processamento da característica de forma

3.4 Processamento das características de som

A classificação e a segmentação de áudio podem ser utilizadas para criar ferramentas para administração de conteúdo, onde um clipe de áudio pode ser automaticamente classificado e organizado em um banco de dados.

A percepção auditiva humana é complexa, por esse motivo, extrair características descritivas do áudio se torna uma tarefa difícil. Existem inúmeras características que podem ser extraídas do áudio, entretanto nenhuma característica que tivesse 100% de precisão para distinguir classes diferentes foi encontrada na literatura, no entanto, é possível alcançar altas precisões combinando várias características, pois cada uma delas possui uma tarefa que pode ou não ser mais apropriada para um determinado fim, dentro da classificação do áudio.

Um *frame* é definido como um grupo de amostras próximas que duram, mais ou menos, de 10 a 40m dentro das quais é possível supor que o sinal de áudio está estável e podem ser extraídas características de curto tempo como

o volume e o coeficiente da Transforma de Fourier. O conceito de *frame* de áudio vem de um processo tradicional de sinal da fala cuja análise de um intervalo muito curto é considerada mais apropriada [LI01].

Há muitas características que podem ser usadas para classificar segmentos de áudio, geralmente elas são extraídas em dois níveis:

(i) frames de curta duração e;

(ii) clipes de longa duração

Para uma característica revelar o significado semântico de um sinal de áudio, é preciso realizar uma análise durante um período maior de tempo, geralmente de 1seg. a 10seg. Esse intervalo de clipe de áudio será chamado nesta dissertação de janela. Um clipe de áudio consiste em uma sequência de *frames*.

A maioria das caracaterísticas em nível de *frame* foram herdadas de processos tradicionais dos sinais da fala. Em geral, podem ser separadas em duas categorias:

(i) características no domínio do tempo, que são computadas diretamente das ondas de áudio;

(ii) características no domínio da freqüência, que são derivadas da transformada de Fourier, de amostras em um *frame*.

Após a leitura de [SCHEIRER97], [FOOTE99] e [TONG01], e realização de alguns testes de implementação, no sentido de escolher um melhor agrupamento para obtenção de melhores resultados, as seguintes características de áudio foram utilizadas neste trabalho:

- Taxa de cruzamento de zero
- Magnitude do espectro delta
- Raiz quadrada média do somatório das intensidades do sinal
- Energia de curto tempo

- Razão dos valores de características altas
- Razão dos valores de características baixas
- Razão da energia de banda

3.4.1 Taxa de Cruzamento de Zero (ZCR - Zero Crossing Rate)

A taxa de cruzamento de zero é uma das medidas mais robustas e indicativas para diferenciar sons com e sem voz. Também é amplamente usada na classificação de música ou fala. [SCHEIRER97] usa o ZCR para classificar o áudio entre música e fala, [TZANETAKIS02] o usa para classificar o áudio em gêneros diferentes de música e [GOUYON00] usa o ZCR para classificar sons de percussão.

Consiste em um método de análise no domínio do tempo, que corresponde à medida da quantidade de mudanças da amplitude do sinal de voz positivo para negativo e vice-versa. Se a taxa é baixa, indica um segmento sem voz do sinal, caso a taxa de cruzamento de zero seja alta, indica grande probabilidade do segmento conter voz. A taxa de cruzamento de zeros também é maior para ambientes barulhentos se comparados à ambientes só com voz.

O ZCR é definido pelo número de cruzamentos de zeros no domínio do tempo em uma janela de processamento, como apresentado na equação 3.5, onde *sign* é 1 para argumentos positivos e 0 para argumentos negativos, M é o número total de amostras em uma janela de processamento e $x(m)$ é o valor da amostra m.

$$ZCR = \frac{1}{M-1} \sum_{m=0}^{M-1} |sign(x(m)) - sign(x(m-1))| \tag{3.5}$$

A taxa de cruzamento de zero está correlacionada ao conteúdo de freqüência de um sinal, por isso os sons com voz e sem voz têm

respectivamente alta e baixa taxa de cruzamento de zero. Isto resulta em uma alta variação de ZCR. A música não tem essa variação em ZCR, mas deve ser dito que algumas partes da música têm variações parecidas.

A implementação desta equação gera um algoritmo simples com uma complexidade computacional baixa.

3.4.2 Magnitude do Espectro Delta (DSM – Delta Spectrum Magnitude)

[TZANETAKIS02] e [LU02] usam a Magnitude do Espectro Delta para diferenciar fala e musica, e é usada, também, para diferenciar fala e sons musicais. O espectro da freqüência, pode ser usado nessa característica e em várias outras a partir da Transformada Discreta de Fourier (TDF) do sinal, como na equação 3.6. O espectro de um *frame* de áudio se refere à transformada de Fourier das amostras nesse *frame*. A dificuldade de usar apenas o espectro como uma característica em nível de *frame* está em sua dimensão muito grande. Pode-se constatar isto através da equação 3.6, onde k é a freqüência no *frame* n, $x(m)$ é a entrada do sinal, $w(m)$ é a função da janela e L é o comprimento da janela.

$$A(n,k)=\left| \sum_{m=-\infty}^{\infty} x(m)w(nL-M)e^{-j\left(\frac{2\pi}{L}\right)km} \right|$$

(3.6)

A definição matemática da Magnitude do Espectro Delta é dada pela Equação 3.7, onde *N* é o número total de *frames, K* é a ordem da TDF, δ é um valor suficientemente pequeno de maneira a evitar *overflow,* e $A(n,k)$ é a Transformada Discreta de Fourier do frame *n*.

$$DSM = \frac{1}{(N-1)(K-1)} \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} \left[\log(A(n, k\delta) - \log(A(n-1, k) + \delta) \right]^2$$

(3.7)

3.4.3 Raiz Quadrada Média do Somatório da Intensidades do Sinal (RMS – Root Mean Square)

O valor da Raiz Quadrada Média é uma medida da energia em um sinal. A variação do valor de RMS pode ser decisivo para a discriminação entre fala e música. O valor é definido como a raiz quadrada da média do somatório das intensidades do sinal ao quadrado, como na Equação 3.8, onde M é o número total de amostras em uma janela de processamento e $x(m)$ é a intensidade do sinal na amostra m (domínio do tempo).

$$RMS = \sqrt{\frac{1}{M} \sum_{m=0}^{M-1} x^2(m)}$$

(3.8)

3.4.4 Energia de Curto Tempo (STE – Short Time Average Energy)

A Energia de Curto Tempo é uma característica simples e amplamente utilizada em vários esquemas de classificação. [LI01] e [LU02] usam STE para classificar e segmentar o áudio.

O STE é definido como a soma de uma seqüência de dados no domínio do tempo elevado ao quadrado, como na Equação 3.9, onde M é o número total de amostras na janela de processamento e $x(m)$ é o valor da amostra m.

$$STE = \sum_{m=0}^{M-1} x^2(m)$$

(3.9)

A Energia de Curto Tempo discrimina trechos de sinal contidos da voz como também daqueles trechos que contém silêncio. O importante neste caso é a escolha adequada da janela assim como do intervalo a ser analisado (comprimento da janela), pois se for pequeno demais as características se perdem lentamente.

3.4.5 Razão dos Valores de Características Altos *(HFVR – HighFeature-Value Ratio)*

[TONG01] propõe uma variação da taxa de cruzamento de zero que diferencia, principalmente a fala da música. Essa característica é chamada de Razão dos Valores de Características Altos e está definida matematicamente como na Equação 3.10, onde N é o número total de *frames*, n é o índice do frame, $ZCR(n)$ é a taxa de cruzamento de zeros no *frame* n, $avZCR$ é a média ZCR em uma janela de processamento de um segundo e $sin(.)$ é a função seno.

$$HFVR = \frac{1}{2\mathrm{N}} \sum_{n=0}^{N-1} [\sin(ZCR(n) - 1{,}5\, avZCR) + 1] \tag{3.10}$$

3.4.6 Razão dos Valores de Características Baixos (LFVR – Low Feature-Value Ratio)

Similar ao HFVR, [TONG01] propõe uma variação da característica de Energia de Curto Tempo usada para diferenciar voz e silêncio e sons ambientais e silêncio. É matematicamente definida como na Equação 3.11, onde N é o número total de *frames*, n é o índice do *frame*, $STE(n)$ é a energia de curto tempo no frame n, $avSTE$ é a média STE na janela de processamento e $sin(.)$ é a função.

$$LFVR = \frac{1}{2\mathrm{N}} \sum_{n=0}^{N-1} [\sin(0{,}5\, avSTE - STE(n) + 1]$$

(3.11)

3.4.7 Razão da Energia de Banda (BER – Band Energy Ratio)

A Razão da Energia de Banda pode ser definida de várias formas, se refere à razão de energia em determinada banda de freqüência no total de energia. O espectro da freqüência é dividido em quatro sub-bandas. A BER é calculada dividindo cada energia de banda pelo total de energia e é calculada pela Equação 3.12, onde $h = M/4$ para os experimentos conforme [LI01].

$$BER = \frac{\sum_{n}^{h} f_i(n)}{\sum_{n=0}^{N} f_i(n)}$$

(3.12)

3.5 Medidas de Similaridade

Os sistemas de recuperação de informação de texto adotam palavras-chaves para indexarem os documentos e também para traduzirem a necessidade de informação do usuário a uma consulta. A representação de documentos e consultas por palavras-chave é uma simplificação, pois parte da semântica dos documentos podem ser perdidas quando é feita a transformação em um conjunto de palavras-chaves. O mesmo ocorre quando a recuperação é de informações de vídeo. Ao se transformar o conteúdo do vídeo em um vetor de características, parte do conteúdo semântico do mesmo também é perdido. Outra simplificação que ocorre nos dois sistemas também faz com que informações sejam perdidas nessas transformações: as ocorrências de palavras-chave (em recuperação textual) e características (em recuperação de vídeos) são consideradas independentes; ou seja, conhecer alguma informação sobre a palavra ou característica i nada implica sobre seu vizinho $i + 1$.

Quando se tem um espaço vetorial e deseja-se medir a distância entre 2 vetores - e conseqüentemente entre 2 vetores de características - pode-se utilizar diversas métricas. Neste trabalho, será utilizado o modelo vetorial descrito a seguir.

3.5.1 Modelo Vetorial

Uma medida de similaridade que considera a distância entre dois vetores é o chamado Modelo Vetorial. Este modelo utiliza um espaço v-dimensional, ou seja, $\vec{q} = (c_{1q}; c_{2q}; : : : ; c_{vq})$ e $\vec{p} = (c_{1p} ; c_{2p} ;... ; c_{vp})$, onde $\vec{q}$ é o vetor de características extraídas das cenas do vídeo obtidas manualmente, e $\vec{p}$ é o vetor de características das cenas, no banco de dados, e c_{ij} é a característica i associada a $i=q$ ou $i=p$. Operações algébricas são efetuadas sobre tais vetores para obter-se uma ordenação dos documentos (cenas recuperadas) [YATES99].

A similaridade entre a consulta e cada documento é calculada, neste trabalho, pela correlação entre os vetores $\vec{q}$ e $\vec{p}$. Essa correlação é o cosseno do ângulo formado entre os dois vetores, dado pela Equação 3.13, onde $|p|$ e $|q|$ são as normas (quadrática) da classificação manual e automática.

$$sim(\vec{p},\vec{q}) = \frac{p^t q}{|p||q|} = \frac{\sum_{i=1}^{n} c_{ip} \times c_{iq}}{\sqrt{\sum_{i=1}^{v} c_{ip}^2} \times \sqrt{\sum_{i=1}^{v} c_{iq}^2}}$$

(3.13)

A norma da classificação manual não influencia o resultado da ordenação e a norma da automática permite a normalização no espaço dos documentos.

As principais vantagens do modelo vetorial são: possibilidade de utilização de pesos para os termos (características) melhorando o desempenho dos sistemas de recuperação de cenas; o casamento parcial entre a classificação e as cenas obtidas; e a ordenação dos documentos na classificação pela similaridade parcial com a manual.

A principal desvantagem é considerar, como premissa, a independência dos termos. Esta premissa é simplificadora, mas permite que os cálculos sejam funções lineares.

Neste trabalho, adotou-se o modelo vetorial para medir a similaridade entre dois blocos de quadros f_i e $f_{i,}$ como $P(f_i|f_j)$. Um bloco é o agrupamento de 30 quadros, devido ao fato de, em um segundo, serem transmitidos 30 quadros.

3.6 Teoria Bayesiana

Redes bayesianas[3] são tipos específicos de redes de conhecimento. Representa uma distribuição de probabilidade conjunta, utilizando fundamentos matemáticos da probabilidade e estatística em conjunto com a teoria dos grafos presente na computação.

A vantagem principal das redes bayesianas, quando comparadas com outras representações de probabilidades, está no fato de poderem representar relacionamentos probabilísticos de uma forma concisa. O mecanismo de inferência é baseado no conceito de probabilidade condicional e no teorema de Bayes. A computação de probabilidade, embora seja exponencial no pior caso, é eficiente em várias situações práticas [RODRIGUES03].

Formalmente, uma rede bayesiana consiste de [JENSEN01]:

(i) Um conjunto de variáveis e um conjunto de linhas direcionadas (vetores) entre essas variáveis;

(ii) Cada variável tem um conjunto finito de estados mutuamente exclusivos;

(iii) As variáveis, juntamente com os arcos, formam um grafo direcionado acíclico (GDA) (um grafo direcionado é acíclico se, e somente se, existe apenas um único caminho entre duas variáveis distintas quaisquer);

(iv) Para cada variável A, com pais $B_1,B_2, \ldots ,B_n$, está associada uma tabela de potenciais $P(A|B_1,B_2, \ldots ,B_n)$.

---

[3] Assim denominada por ter origem no teorema que leva o nome do reverendo Thomas Bayes, desenvolvido por Simon de La Place em 1812.

Naturalmente, na escolha do método consideram-se critérios diferentes, como a facilidade, o custo e a demora na implantação de uma solução. Segundo [NAIM04] os seguintes aspectos das redes bayesianas as fazem

preferíveis a outros modelos:

• Aquisição de conhecimentos. A possibilidade de juntar e fundir conhecimentos de naturezas diversas num mesmo modelo: dados históricos ou empíricos, experiência (expressa na forma de regras lógicas, de equações, de estatísticas ou de probabilidades subjetivas), observações;

• Representação de conhecimentos. A representação gráfica de uma rede bayesiana é explícita, intuitiva e compreensível para uma pessoa não especialista, o que por sua vez, facilita a validação do modelo, suas evoluções eventuais e, sobretudo a sua utilização. Tipicamente, um decisor tem mais confiança sobre um modelo no qual ele compreende o funcionamento, do que num modelo tipo “caixa preta”;

• Utilização de conhecimentos. Uma rede bayesiana é multifuncional: pode-se usar o mesmo modelo para avaliar, prever, diagnosticar, ou otimizar as decisões, o que contribui para rentabilizar o esforço gasto na construção da rede bayesiana;

• Qualidade da oferta com relação aos programas. Hoje existem inúmeros programas para aproveitar e tratar as redes bayesianas. Estas ferramentas apresentam funcionalidades mais ou menos evoluídas: aprendizagem de probabilidades, aprendizagem da estrutura da rede bayesiana, possibilidade de integrar variáveis contínuas, variáveis de utilidade e de decisão dentre outras.

As redes bayesianas também permitem analisar grandes quantidades de dados para serem extraídos conhecimentos úteis em tomada de decisões; para controlarem ou preverem o comportamento de um sistema; para diagnosticarem as causas de um fenômeno etc. e, também, são utilizadas em várias áreas: na saúde (diagnóstico, localização de genes); na indústria (controle de autômatos ou de robôs); em computação e redes (agentes inteligentes); em marketing (mineração de dados, gestão da relação com os clientes); em banco e finanças (análise financeira) e na gestão (tomada de decisões, gestão de conhecimento e risco).

De acordo com o tipo de aplicação, a utilização prática de uma rede bayesiana pode ser considerada da mesma maneira que modelos como: redes neurais, sistemas especialistas, árvores de decisão, modelos para análise de dados (regressão linear), modelos lógicos etc.

3.6.1 Distribuição Conjunta Probabilística

Fenômenos de observação necessitam de modelos matemáticos (determinísticos ou probabilísticos) que simplifiquem seu entendimento.

A fim de verificar a validade de um modelo na representação de um fenômeno, é preciso deduzir um certo número de conseqüências do fenômeno, e, a seguir, comparar esses resultados previstos com as observações. É muito importante distinguir a própria observação do fenômeno do modelo que o representa. Por isso, é necessária a utilização de modelos apropriados à circunstância dos experimentos. Nesse sentido, há modelos determinísticos e não determinísticos (probabilísticos).

Nos modelos determinísticos, os resultados do experimento são sempre o mesmo [FONSECA96]. Já os probabilísticos são modelos matemáticos para experimentos aleatórios, isto é, experimentos que, repetidos sobre as mesmas condições, produzem resultados distintos.

Dados dois eventos, A e B, a probabilidade condicionada do evento A quando B tiver ocorrido, denotado por $P(A|B)$ é dada pela Equação 3.14:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{com} \quad P(B) \neq 0 \tag{3.14}$$

Seja o experimento E: "*lançar um dado*" e o evento A = *{face superior é número 3}*. Então $P(A) = 1/6$. Considerando agora outro evento B = *{sai um número ímpar}* = *{1, 3, 5}*.

Pode haver interesse em avaliar a probabilidade do evento A condicionada à ocorrência do evento B. Em símbolos esta probabilidade é designada por $P(A|B)$; lê-se: "probabilidade de A dado que B ocorreu".

Assim: $P(A|B) = 1/3$.

Dada a informação da ocorrência de um evento, haverá a redução do espaço amostral; no exemplo $S = \{1, 2, 3, 4, 5, 6\}$ foi reduzido para $S^* = \{1, 3, 5\}$ e é neste espaço amostral reduzido que será avaliado a probabilidade do evento A seguir a B.

Um evento A é considerado independente de um outro evento B se a probabilidade de A é igual à probabilidade condicional de A dado B, isto é, se $P(A) = P(A|B)$.

Evidente que, se A é independente de B, B é independente de A; assim: $P(B) = P(B \mid A)$ Considerando o teorema do produto [LIPSCHUTZ68], pode-se afirmar que se A e B são independentes, então: $P(A \cap B) = P(A)\ P(B)$

Dados "n" eventos $A1, A2, \ldots, An$, diz-se que eles são independentes se eles forem independentes 2 a 2; 3 a 3; ...; n a n, ou seja, as igualdades abaixo forem verificadas:

$$P(A_1 \cap A_2) = P(A_1)P(A_2); \ldots; P(A_{n-1} \cap A_n) = P(A_{n-1})P(A_n)$$
$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3); \ldots; P(A_{n-2} \cap A_{n-1} \cap A_n) = P(A_{n-2})P(A_{n-1})P(A_n)$$
$$\vdots$$
$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2)P(A_3)\ldots P(A_{n-1})P(A_n)$$

3.6.2 Teorema de Bayes

Sejam $A_1, A_2, A_3, \ldots, A_n$, eventos mutuamente exclusivos tais que $A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_n = S$. Sejam $P(A_i)$ as probabilidades conhecidas dos vários

eventos, e B um evento qualquer de S tal que se sabe todas as probabilidades condicionais $P(B|A_i)$.

Então para cada "i":

$$P(A_i|B)=\frac{P(A_i)P(B|A_i)}{(P(A_1)P(B|A_1))+(P(A_2)P(B|A_2))+\ldots+(P(A_n)P(B|A_n))}$$

(3.15)

O resultado acima é importante, pois relaciona probabilidades "*a priori*" $P(A_i)$ com probabilidades "*a posteriori*" $P(A_i|B)$, isto é, probabilidade de A_i depois que B ocorrer. Não depende de repetições de experimentos, permitindo a representação de conhecimentos incertos através da distribuição de probabilidade conjunta (DPC) obtida pelo produto de distribuições condicionadas. $P(A|B)$ representa a probabilidade do evento A (hipótese) condicionada a ocorrência de um outro evento B (evidência).

O teorema de Bayes pode ser facilmente reescrito na forma que expressa como obter a probabilidade posterior de uma hipótese A, depois da observação de uma evidência B, dada a probabilidade, "*a priori*", de A e a possibilidade de observar B existindo A ser o caso:

$$p(A|B)=\frac{p(B|A)\,p(A)}{p(B)}$$

(3.16)

Na forma geral, o teorema de Bayes pode ser usado como apoio no caso onde a hipótese é uma proporção no domínio de conhecimento e a evidência é observada por algumas condições. Entretanto, pode também apoiar o caso onde uma hipótese é um parâmetro em um modelo de conhecimento tendo um certo valor (ou valores de distribuição) ou que o modelo tem uma certa estrutura, e a evidência é qualquer aproximação nos dados ocorridos.

3.6.3 Teoria dos Grafos

Um grafo G consiste de dois conjuntos V e E. O conjunto V é um conjunto de nós finito e não vazio. O conjunto E é um conjunto de pares de nós (nós); estes pares são chamados de arestas (arcos). A notação $V(G)$ e $E(G)$ representa os conjuntos de nós e arestas, respectivamente, do grafo G. Escreve-se também $G = (V,E)$ para representar um grafo. Em um grafo não direcionado, o par de nós representa qualquer arco sem direção. Assim, os pares (u,v) e (v,u) representam a mesma aresta. Em um grafo direcionado, cada arco é representado por um par ordenado $<u,v>$; u é a cauda e v a cabeça do arco. Então $<u,v>$ e $<v,u>$ representam dois arcos diferentes. Exemplos de grafos são mostrados na Figura 3.7:

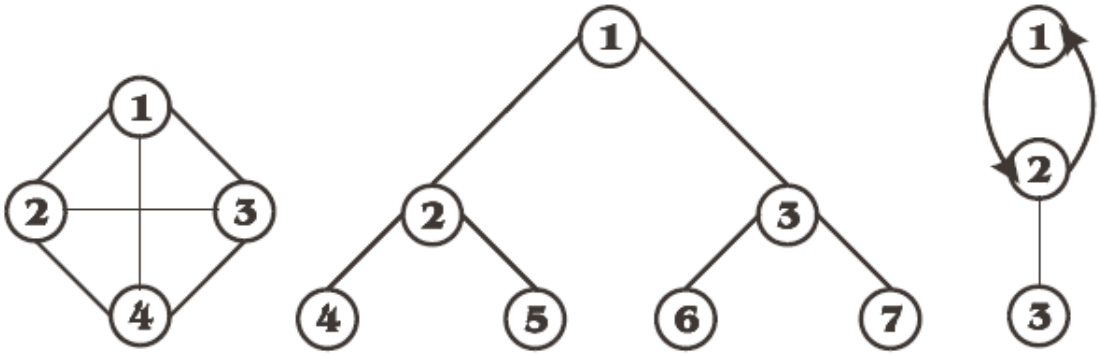


Figura 3.7: Exemplos de Grafos

Definidos os arcos e nós de um grafo como conjunto, apresentam-se as restrições:

a) Um grafo não deve ter um arco de um nó v que volta para ele mesmo. Ou seja, os arcos (v,v) e $<v,v>$ não são válidos.

b) Um grafo não deve ter múltiplas ocorrências de um mesmo arco. Se permitisse esta restrição obter-se-ia um grafo múltiplo.

O número de pares não dirigidos distintos (u,v) com u v em um grafo com

n nós é $n(n-1)/2$. Isto é o número máximo de arcos em qualquer n-nó, em grafos não dirigidos.

Em um arco (u,v) em $E(G)$ é dito que os nós u ,v são adjacentes e o arco (u,v) é incidente sobre os nós u e v.

O grau de um nó é o número de arcos incidentes neste nó. Se d_i é o grau do nó i em um grafo G com n nós e e arcos, então o número de arcos é dado pela Equação 3.17::

$$e = \left(\sum_{i=1}^{n} d_i \right) / 2 \tag{3.17}$$

Existem duas formas de representar um grafo $G = (V,E)$, com uma coleção de lista de adjacência ou uma matriz de adjacência. A lista de adjacência é a representação normalmente utilizada, porque ela fornece uma forma compacta para representar grafos esparsos, os quais $|E|$ é muito menor do que $|V2|$. A representação matriz de adjacência, porém, pode ser preferida quando o grafo é denso, $|E|$ é próximo de $|V2|$, ou quando é preciso capacidade de responder rapidamente se um arco conecta dois nós informados. Melhores detalhes sobre introdução e representações de grafos veja [HOROWITZ97, BALAKRISHNAN97].

3.6.3.1. Grafos Acíclicos Direcionados

Formalmente, as redes de conhecimento bayesianas são grafos direcionados acíclicos nos quais os nós representam variáveis aleatórias com medidas de incerteza associadas. Os arcos significam a existência de uma influência causal direta entre as variáveis conectadas, e a força destas influências é quantificada por probabilidades condicionais [PEARL88].

Para que se possa verificar se um grafo direcionado e acíclico representa uma rede bayesiana, existe uma condição necessária e suficiente: cada variável X (nó X) do grafo deve ser condicionalmente independente de todos os nós que não são seus descendentes exceto seus pais, ou seja, se os valores dos nós diretamente conectados a uma variável são conhecidos, todos os outros nós do grafo são irrelevantes na definição do valor da variável em questão. Baseado nesta condição, pode-se formular um procedimento recursivo para a construção de uma rede bayesiana. Dada uma distribuição de probabilidade conjunta $P(x_1, ..., x_n)$ e uma determinada ordem destas variáveis dada por d, inicia-se a construção do grafo escolhendo o nó raiz (X_1) e atribuindo-se a ele a probabilidade marginal $P(x_1)$. Em seguida, acrescenta-se mais um nó (X_2) no grafo. Caso X_2 seja dependente de X_1, X_2 tem X_1 como pai, então se traça um arco direcionado que vai de X_1 a X_2 e quantifica-se este arco com $P(x_2/x_1)$, caso contrário, mantém-se X_1 e X_2 desconectados e atribui-se uma probabilidade prévia (*a priori*) $P(x_2)$ a X_2. Ao se atingir o i-ésimo nó (X_i), traçam-se os arcos de X_i a todos os seus pais (Π_i) e quantifica-se este grupo de arcos com $P(x_i/\Pi_i)$ e assim obtém-se a rede bayesiana que representa todas as dependências que são conseqüências da definição dos pais das variáveis [PEARL88].

A rede bayesiana da Figura 3.8 [PEARL88] mostra a relação causal direta existente entre suas variáveis ligadas por arcos orientados. Observa-se que x_i causa x_j se houver um arco de x_i para x_j. A relação causal tem as seguintes possibilidades:

(i) uma variável pode causar uma ou mais variáveis (filhas);

(ii) uma variável pode sofrer a influência causal de uma ou mais variáveis (pais).

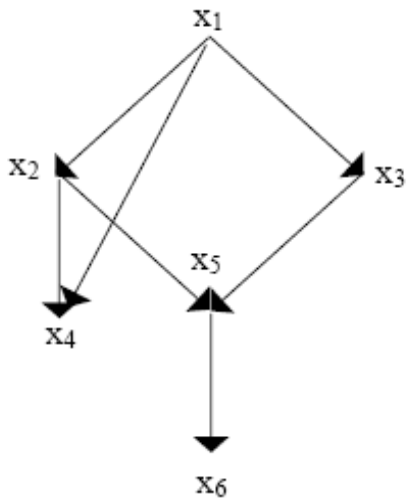


Figura 3.8 Rede Bayesiana [PEARL88]

A quantificação da relação causal é atualizada por uma distribuição condicional, a qual condiciona a variável causada à(s) sua(s) causadora(s).

Exemplos:

$P(x_2/x_1)$,

$P(x_5/x_2,x_3)$.

Representa-se a distribuição conjunta como o produto das distribuições condicionais dadas pelas relações causais na rede, assim:

$P(x_1, x_2, ..., x_6) = P(x_6/x_5)\ P(x_5/x_2, x_3)\ P(x_4/x_1, x_2)\ P(x_3/x_1)\ P(x_2/x_1)\ P(x_1)$.

3.6.4 Inferência em Redes Bayesianas

Tendo-se uma rede bayesiana definida, podem-se extrair os conhecimentos nela representados através de um processo chamado inferência. Existem vários métodos para a realização de inferência em uma rede bayesiana [HRUSCHKA97].

Sejam X e Y variáveis aleatórias representadas por nós da rede bayesiana, e sejam x e y seus respectivos valores. Se o valor da variável X é causa direta de influência no valor da variável Y, um vetor de X para Y é inserido no grafo (X é dito pai de Y). A intensidade do relacionamento entre X e Y é expressa pela probabilidade condicional $P(y|x)$. Diz-se, probabilidade de $Y = y$, caso $X = x$.

Seja B uma rede bayesiana, seja Y uma variável aleatória em B, seja y o valor de Y, seja P_Y o conjunto de variáveis pais de Y e seja p_y o conjunto dos valores de P_Y, como indicado na Figura 3.8. A influência de P_Y sobre Y pode ser especificada por uma função F que satisfaça a Equação 3.18:

$$\sum_{\forall y} F(y, Py) = 1 \qquad 0 \leq F(y, Py) \leq 1 \tag{3.18}$$

A função $F(y, p_y)$ fornece uma quantificação da probabilidade condicional $P(y|p_y)$. Esta especificação representa a distribuição conjunta de probabilidade

para os nós da rede B [PEARL88].

Para exemplificar, considere a Figura 3.9, que representa a distribuição conjunta de probabilidade $P(x1, x2, x3, x4, x5)$ para as variáveis aleatórias $\{X1,X2,X3,X4,X5\}$, onde $\{x1, x2, x3, x4, x5\}$ são seus respectivos valores. O nó X_1 é denominado raiz da rede (nó sem pai) e é o pai de X_2 e X_3. A probabilidade $P(x_1)$, associada com o valor x_1 do nó raiz X_1, é denominada probabilidade *a priori* e é utilizada para representar o conhecimento prévio sobre o domínio modelado. Dado o valor da variável X_1, as variáveis X_2 e X_3 são independentes. Dados os valores das variáveis X_2 e X_3, as variáveis X_4 e X_5 são independentes. Devido a essas independências, a distribuição conjunta de probabilidade $P(x_1, x_2, x_3, x_4, x_5)$ pode ser calculada pela Equação 3.19:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)\, P(x_2|x_1) P(x_3|x_1)\, P(x_4|x_2, x_3)\, P(x_5|x_3) \tag{3.19}$$

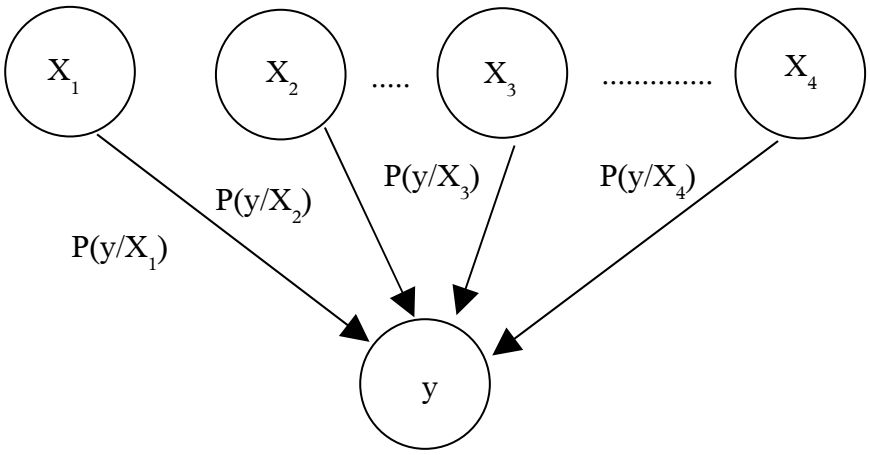


Figura 3.9: Exemplo de rede Bayesiana.

A Figura 3.10, embora represente uma situação fictícia e improvável,

ilustra bem os conceitos mencionados acima. Nela, pretende-se estabelecer a influência causal da variável Fraude (cartão fraudado), Idade e Sexo sobre compras de Gasolina e Jóias.

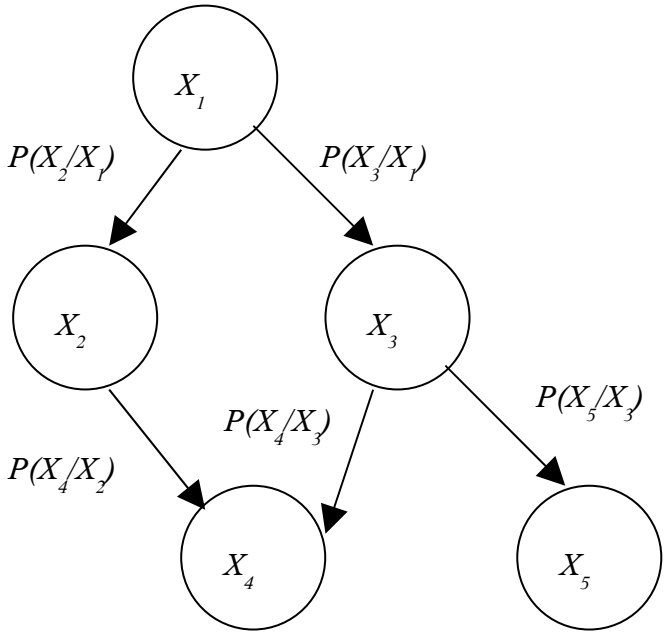


Figura 3.10: Rede Bayesiana com independências declaradas.

O conjunto de variáveis $X = \{Fraude, Idade, Sexo, Gasolina, Joias\}$ retrata as variáveis do modelo; os círculos representam tanto os nós da rede quanto as variáveis do conjunto X; os vetores representam o relacionamento causal entre as variáveis; e os parâmetros numéricos são representados pelas distribuições

marginais ou condicionais do modelo e que são dadas nas Tabelas 3.3, 3.4, 3.5, 3.6 e 3.7 (lembrando que a Fraude influencia a compra de Gasolina, e que Fraude, Idade e Sexo, conjuntamente, influenciam a compra de Jóias). A distribuição conjunta das variáveis do modelo pode ser expressa como (Figura 3.11):

$$P(Fraude, Idade, Sexo, Gasolina, Joias) = P(Fraude)\,P(Idade)\,P(Sexo)\,P(Gasolina \mid Fraude)\,P(Jóias \mid Fraude, Idade, Sexo)$$

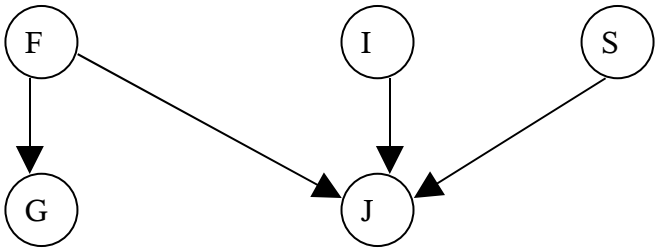


Figura 3.11: Estrutura de rede bayesiana para detecção de problemas de fraude:
F = Fraude; I = Idade; S = Sexo; G = Gasolina; J = Jóias.

Tabela 3.3: P(F)

P(F=Falso)	*P(F=Verdadeiro)*
0.99999	*0.00001*

Tabela 3.4: P(I)

P(I=<30)	*P(I=30 – 50)*	*P(I=>50)*
0.25	*0.40*	*0.35*

Tabela 3.5: *P(S)*

P(S=Feminino)	P(S=Masculino)
0.5	0.5

Tabela 3.6: *P(G|F)*

F	*I*	*S*	*P(J=Falso)*	*P(J=Verdadeiro)*
F	*<30*	*F*	*0.9995*	*0.0005*
V	*<30*	*F*	*0.95*	*0.05*
F	*<30*	*M*	*0.9999*	*0.0001*
V	*<30*	*M*	*0.95*	*0.05*
F	*30 - 50*	*F*	*0.998*	*0.002*
V	*30 - 50*	*F*	*0.95*	*0.05*
F	*30 - 50*	*M*	*0.9996*	*0.0004*
V	*30 - 50*	*M*	*0.95*	*0.05*
F	*>50*	*F*	*0.999*	*0.001*
V	*>50*	*F*	*0.95*	*0.05*
F	*>50*	*M*	*0.9998*	*0.0002*
V	*>50*	*M*	*0.95*	*0.5*

Tabela 3.7: P(J|F, I, S).

F	*P(G=Falso)*	*P(G=Verdadeiro)*
F	*0.99*	*0.01*
V	*0.8*	*0.2*

Suponha que para o problema de detecção de fraude queira se conhecer a probabilidade de fraude dadas as observações das outras variáveis. Esta probabilidade não está armazenada diretamente no modelo e é necessário calculá-la. O cálculo das probabilidades pode ser realizado mediante uma simplificação e fatorização de variáveis:

$$P(F|I,S,G,J)=\frac{P(F,I,S,G,J)}{P(I,S,G,J)}=\frac{P(F,I,S,G,J)}{\sum_F P(F=Falso,I,S,G,J)}$$

Dadas as independências condicionais, a equação acima pode ser expressa como:

$$P(F|I,S,G,J)=\frac{P(F)P(I)P(S)P(G|F)P(J|F,I,S)}{\sum_F P(F=Falso)P(I)P(S)P(G|F=Falso)P(J|F=Falso,I,S)}$$

$$P(F|I,S,G,J)=\frac{P(F)P(G|F)P(J|F,I,S)}{\sum_F P(F=Falso)P(G|F=Falso)P(J|F=Falso,I,S)}$$

A partir desta simplificação, substituem-se os valores das probabilidades e eliminam-se, seguindo uma ordem específica, as variáveis *I, S, G, J*.

A tarefa fundamental para qualquer sistema de inferência probabilística é calcular a distribuição de probabilidade *a posteriori* para um conjunto de variáveis, dado o valor exato de alguma outra variável. Em [RUSSELL95] são denominadas as primeiras "variáveis consulta" e as últimas "variáveis evidência". Deste modo o sistema calcula *P(Consulta|Evidencia)*.

Redes bayesianas são flexíveis a ponto de qualquer nó poder servir como consulta ou como evidência. Uma vez construída uma representação probabilística através do modelo e redes bayesianas, para incerteza presente no relacionamento entre variáveis de um domínio de dados, uma das tarefas mais importantes está em obter estimativas de probabilidades de eventos relacionados aos dados, à medida que novas informações ou evidências sejam conhecidas.

A inferência em redes bayesianas, mediante o cálculo das probabilidades *a posteriori*, permite responder a uma série de consultas ou *queries* sobre um domínio modelado de dados através das redes bayesianas, a partir de nova informação (evidência) conhecida.

Todas as técnicas vistas neste capítulo foram utilizadas nos experimentos deste trabalho que serão apresentados no capítulo 4, em Metodologia.

4. Metodologia Usada nos Experimentos

A metodologia de testes usada nessa dissertação consiste em adquirir alguns arquivos de áudio como noticiário, filme, desenho, show e depois de um pré-processamento, tratar as características audiovisuais para detectar as transições de cena.

4.1 Pré-processamento

Cinco vídeos, totalizando 25 minutos, foram selecionados, para os experimentos, que serão apresentados no Capítulo 5. Nestes vídeos há dois noticiários, sendo um com duração de 6 minutos e 30 segundos; e outro com duração de 5 minutos e 45 segundos. Esses noticiários mostram um telejornal da cidade de Petrópolis com entrevistas, comerciais e diversas reportagens. Foram escolhidos pela grande variedade de recursos sonoros e visuais. Dentre os sonoros observa-se diversas pessoas falando ao mesmo tempo; apenas o apresentador; pessoas falando com ruído entre outros. As variedades visuais incluem diversas imagens com muitas tonalidades de cores e formas bem diferentes em diversos quadros e cenas. Um dos vídeos consiste de um show com duração de 2 minutos e 45 segundos. Este foi selecionado pelo fato de conter enorme variedade de recursos sonoros tais como: música com ruído; pessoas falando com fundo musical; várias pessoas falando; apenas uma pessoa falando, além de diversos recursos visuais. O quarto vídeo é um desenho animado com duração de 5 minutos; selecionado para uma comparação entre vídeos com cenas reais e vídeos produzidas digitalmente. O quinto vídeo usado para os experimentos é um filme brasileiro com duração de 5 minutos. Este se mostra adequado para análise entre cenas com transições de tomadas abruptas e cenas com transições de tomadas graduais. Estes vídeos fazem um total de 25 minutos de filmes usados para testes.

Após a seleção dos vídeos, apresentados em VHS, os mesmos são transformados em formato .AVI (*Audio-Vídeo Interleaved*) utilizando o software *Adobe Premiere*. Após a digitalização, passam pelo programa *Virtual Dub* onde são separados os quadros do arquivo de som. Nesse momento o áudio é gravado em formato .WAV e os quadros em formato .TIF. A escolha pelas extensões .TIF e WAV se deve ao fato de serem facilmente interpretada pelo software MatLab, onde todos os procedimentos são testados. O software *IrfanView* é usado para renomear todos os quadros a fim de se obter a verdadeira seqüência do vídeo e também para diminuir o tamanho dos quadros, nesse caso para 180 x 120 *pixels*, com tamanho aproximado de 70Kb. A resolução foi reduzida para diminuir o poder computacional de processamento, dado que essa diminuição em nada influenciará o resultado final.

A configuração *default* para exibição de vídeo, no Brasil, é de 30 quadros por segundo, por esse motivo, as imagens .TIF foram particionadas em blocos de 30 imagens/quadros.

O som é particionado em arquivos menores, compostos por 30 segundos de áudio, cada, além disso, foi utilizado o software *Cool Edit* para gerar os arquivos binários, de forma que pudessem ser lidos no MatLab, onde os testes são executados. A divisão do arquivo foi feita para agilizar a etapa de processamento, dado o poder computacional exigido para tal aplicação. Inicialmente, um teste foi realizado utilizando um único arquivo de áudio, com duração de 6 minutos e 30 segundos, o algoritmo funcionou perfeitamente, entretanto o processamento levou aproximadamente 5 horas. Após a divisão, o tempo total de processamento de todos os arquivos, foi de aproximadamente duas horas, utililzando o mesmo computador (Processador *Intel Pentium 4* 3.6GHz HT; 2GB de memória RAM; Sistema Operacional *Windows XP Profissional*), conforme visto na Figura 4.1.

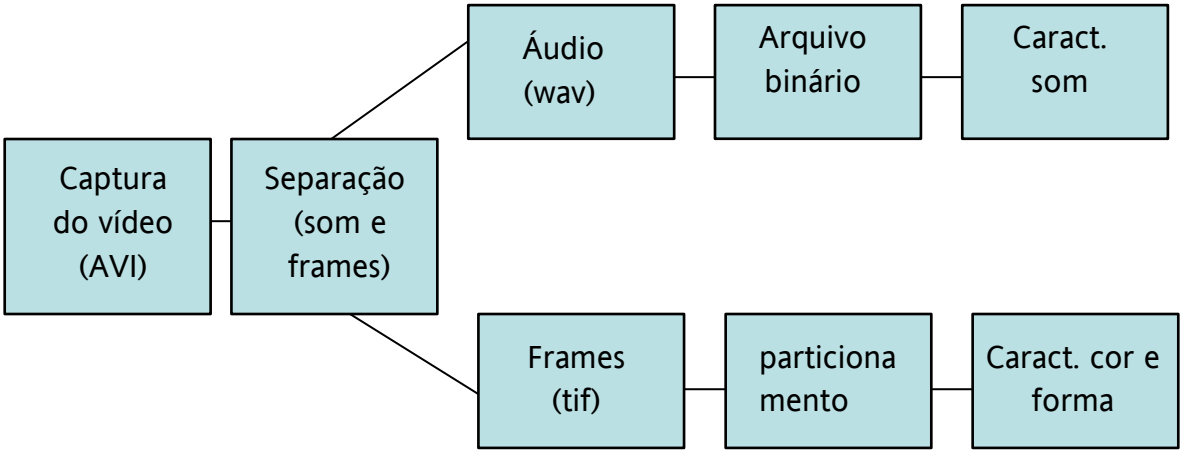


Figura 4.1: Pré-processamento

4.2 Segmentação Manual

Para estabelecer um *ground-truth* para a comparação, o mesmo vídeo foi particionado manualmente, também comparando blocos compostos por 30 quadros, onde cada bloco foi rotulado da seguinte forma: cada função de transição de cena $F(i) \quad (i=1,2,3,\ldots,n)$, onde *n* é o número de blocos, é dada da seguinte forma:

$$F(i) = \begin{cases} 0 & \text{se ocorrer transição na cena;} \\ 1 & \text{caso contrário} \end{cases}$$

ou seja, cada bloco com 30 quadros é comparado a outro bloco, também com 30 quadros; quando esses blocos são similares, considerando as características de cor, forma e som, é atribuído o valor 1 para sua posição no vetor manual; e quando há variações de cor, forma e som entre esses blocos, então é atribuído o valor zero à posição correspondente no vetor.

Neste trabalho, a segmentação manual foi considerada correta e precisa, devido a alguns fatores:

(i) em todos os vídeos, as transições de corte seco ou graduais são bem definidas, portanto fácil de serem identificadas pelo ser humano;

(ii) os vídeos apresentam cenas precisas e foram selecionados principalmente por esta característica, dado que, num filme, por exemplo, identificar cenas manualmente, pode ser uma tarefa difícil;

(iii) duas pessoas treinadas fizeram a análise manual e os resultados se mostraram idênticos.

Os resultados experimentais que serão vistos no Capítulo 5 comparam os resultados gerados pelo sistema, com o vetor produzido manualmente.

4.3 Processamento considerando as características do vídeo (cor, forma e som)

1) Inicialmente o diretório onde se localiza a seqüência de arquivos .TIF do vídeo é informado e também o número de quadros para a formação de um bloco, neste caso, 30 imagens. Além disso, também é informado o diretório onde se encontram os arquivos binários de som. Para cada arquivo de som com 30 segundos, são atribuídos 30 blocos com 30 quadros cada, totalizando 900 quadros;

2) Para cada arquivo de áudio, são calculadas as seguintes características: energia média de tempo curto; razão da energia de banda; magnitude do espectro delta; taxa de cruzamento de zero; raiz quadrada média da soma das intensidades do sinal ao quadrado; razão dos valores de características altas e razão dos valores de características baixas. Cada uma dessas características são normalizadas pelos seus valores máximos e mínimos. Em seguida, é feita a armazenagem em um vetor com 7 posições é armazenado;

3) Os quadros são passados do sistema RGB para o HSV; as 3 bandas são separadas; a imagem em HSV é quantizada e remontada;

4) O cálculo do histograma de cor normalizado no espaço HSV é realizado;

5) Os quadros, no sistema RGB, são convertidos para escala de cinza e, posteriormente para matriz de índices;

6) Os vetores de características de forma são tomados com o histograma do cálculo das tangentes, discretizados em 10 direções diferentes (180º divido por 18º = 10 possíveis direções). Essas direções são calculadas com o filtro Sobel, que é passado na direção X e na direção Y e calculadas as tangentes. As tangentes são histogramadas de 18º em 18º. O histograma é normalizado;

7) O modelo vetorial é usado para calcular a distância entre os quadros. Os quadros são agrupados em blocos de 30 (p e q). Um novo vetor armazena essas distâncias;

8) A teoria das redes bayesianas é utilizada, como detalhada na Seção 4.5, a seguir.

4.4 Utilização do Modelo

O modelo Bayesiano apresentado em [RIBEIRO98, COELHO04, RODRIGUES03, RODRIGUES05] para recuperação de informação possui a forma geral dada pela Equação 4.1:

$$P(A|B)=1-[(1-P_i(A|B))\times(1-P_j(A|B))] \tag{4.1}$$

onde lê-se: $P(A|B)$ é a probabilidade de recuperar o documento A tal

que ocorreu a recuperação do documento B, $P_i(A|B)$ é a probabilidade de ocorrer o documento A tal que ocorreu o documento B com característica i. Assim, $P(A|B)$ é uma combinação linear de várias probabilidades que leva em conta características i. Por exemplo, em [RIBEIRO98], o modelo foi usado para comparar dois documentos textuais, assim $P(A|B)$ é a probabilidade do documento A ser igual ao B considerando que $P_i(A|B)$ é a probabilidade de $A = B$ considerando a freqüência das palavras dos textos A e B, com a característica i.

Em [COELHO04], o modelo foi usado para comparar imagens em um sistema baseado em informações textuais. Assim $P(A|B)$ era a probabilidade da imagem $A = B$ considerando, por exemplo, $P_i(A|B)$ a probabilidade da imagem A ser igual a B levando-se em conta a freqüência de palavras em volta das imagens em uma página HTML, característica i.

Em [RODRIGUES03, RODRIGUES05] esse modelo foi extendido para RIBC como mostra a Equação 4.2:

$$P(A|B)=1-[(1-P_C(A|B))\times(1-P_F(A|B))] \quad (4.2)$$

onde $P(A|B)$ é a probabilidade da imagem $A = B$ e $P_C(A|B)$ e $P_F(A|B)$ são as probabilidades da imagem $A = B$ levando-se em conta características de cor e forma, respectivamente.

Nota-se que esse modelo é facilmente extensível para outras características, apenas incluindo-se novas parcelas de probabilidade na Equação 4.2. Neste trabalho é proposta a inclusão do som nessa equação e $P(A|B)$ é a probabilidade de um bloco de n quadros da cena A ser igual a outro bloco B, do mesmo tamanho. Assim, foi usada a Equação 4.3 para comparar os blocos de quadros.

$$P(A|B)=1-[(1-P_C(A|B))\times(1-P_F(A|B))\times(1-P_S(A|B))] \quad (4.3)$$

onde:

$P(A|B)$ é a probabilidade de um bloco A, de n quadros ser igual a um

bloco B, também de n quadros;

$P_C(A|B)$ é a probabilidade de um bloco A, de n quadros ser igual a um bloco B, também de n quadros, levando-se em conta características de cor;

$P_F(A|B)$ é a probabilidade de um bloco A, de n quadros ser igual a um bloco B, também de n quadros, levando-se em conta características de forma;

$P_S(A|B)$ é a probabilidade de um bloco A, de n quadros ser igual a um bloco B, também de n quadros, levando-se em conta características de som.

Neste trabalho $P_C(A|B)$ é calculado com o modelo vetorial (Seção 3.5.1) entre o histograma HSR-162 de cada um dos quadros do bloco A e cada um dos quadros do bloco B e $P_F(A|B)$ é calculado com o modelo vetorial (Seção 3.5.1) entre o vetor de características de forma de cada um dos quadros do bloco A e de cada um dos quadros do bloco B. Os vetores de características de forma são tomados com o histograma do gradiente da imagem discretizados em 10 direções diferentes (180º divido por 10º) (Seção 3.3.1). Para cada quadro do bloco, tem-se 30 medidas de similaridade entre os quadros do bloco A e do bloco seguinte (B).

Do mesmo modo, $P_S(A|B)$ é calculado com o modelo vetorial (Seção 3.5.1) entre o vetor de características de som do grupo de 30 blocos A e do grupo de 30 blocos de B. Sendo esse vetor de som, composto de 7 características (ou índices), dados por:

1. Energia média de curto tempo (domínio do tempo)
2. Razão da energia de banda (domínio da freqüência)
3. Magnitude do espectro delta (domínio da freqüência)
4. Taxa de cruzamento de zero (domínio do tempo)
5. Raiz quadrada média do somatório das intensidades do sinal ao quadrado (domínio do tempo)
6. Razão dos valores de características altos (domínio do tempo)
7. Razão dos valores de características baixos (domínio do tempo)

As características de forma, cor e som de cada um dos 30 quadros do

bloco são combinadas, usando a Equação 4.3, gerando 30 probabilidades para cada quadro do bloco A e B. A decisão da existência ou não da transição no bloco automaticamente é feita a partir da comparação desses valores com o limiar.

4.5 Explicação dos Experimentos

Após a geração dos arquivos decorrentes do processamento das características, como descrito nas Seções 4.2 e 4.3, o primeiro experimento é realizado comparando o vetor gerado manualmente, com os gerados pelo MatLab. Dessa forma:

1) O vetor de probabilidades é inicializado;
2) Um limiar é estabelecido para comparar as fronteiras, ou seja, até que ponto será considerado uma transição e até que ponto será considerada mesma cena;
3) O vetor L é gerado composto apenas por Os e 1s. Nesse trabalho dois limiares foram utilizados (0.7 e 0.99998), entretanto, o número de acertos considerando o total de vídeos utilizados, foi maior utilizando o limiar 0.99998. Se V for maior que o limiar, há transição de cena, caso contrário, não há.
4) Os dois arquivos (manual (D) e gerado pelo sistema (L)) são comparados, se $D(i) = L(i)$ então o programa acertou, caso contrário, o programa errou;
5) O percentual de acertos é calculado e também o número de falsos positivos e falsos negativos.

5. Resultados e Análises

5.1 Introdução

Neste capítulo, são apresentados os resultados dos experimentos realizados com o modelo bayesiano combinando as seguintes evidências: cor, forma e som.

No intuito de minimizar o erro devido à subjetividade, como já foi explicado anteriormente, os quadros dos vídeos analisados foram classificados manualmente, criando-se um "*ground truth*" para comparação.

A análise manual, utilizando a percepção visual humana, será considerada precisa neste trabalho, devido às transições bruscas ou graduais serem bem perceptíveis/detectadas. Conforme apresentado na Figura 5.1, é possível observar uma transição de corte seco entre os quadros 181 e 182.

Figura 5.1: Transição dos quadros 180, 181 e 182, respectivamente..

A Figura 5.2 mostra outras transições de corte seco entre transição de cena:

Figura 5.2: Transição de corte seco

A Figura 5.3 apresenta quadros aleatórios da mesma cena, que mesclam diferentes objetos, embora sejam mantidas as mesmas tonalidades de cor e o mesmo *background*.

Figura 5.3: Quadros pertencentes à mesma cena

Abaixo, na Figura 5.4 é possível visualizar uma transição gradual entre as cenas.

Figura 5.4: Transição gradual entre cenas

Todos os experimentos e testes foram realizados no LNCC – Laboratório Nacional de Computação Científica e conduzidos em uma máquina com a seguinte configuração:

- Processador *Intel Pentium 4* 3.6GHz HT;
- 2GB de memória RAM;
- Sistema Operacional *Windows XP Profissional*.

5.2 Segmentação Manual

Sendo $g(i)$ a função de transição de cena tomada manualmente ($g(i)=l$ onde i representa o número do bloco; onde $l=0$ quando houver transição de cena e $l=1$ quando não houver transição), os gráficos apresentados a seguir mostram a comparação de ambas ao longo dos blocos.

Para calcular o percentual de acerto, procedem-se da seguinte forma:

Para cada transição, compara-se g e f, e mede-se o número de vezes em que elas coincidem:

Se g(i) corresponde a uma transição então g(i) = 0, se neste *i, f(i) =0*
então acerto = acerto + 1

5.3 Vídeo 1

Duração: 6 minutos e 30 segundos
Cenas: 14
Transições: 13 (definidas manualmente)
Blocos: 390
Quadros: 11.710
Descrição: jornal seguido de comerciais

No gráfico da Figura 5.5 é apresentado o resultado da análise manual, onde *f(i)=0* representa uma transição de cena. No gráfico da Figura 5.6 é possível visualizar as transições de cena automáticas, definidas pelo algoritmo. Quando as Figuras 5.5 e 5.6 são comparadas, são identificadas algumas transições corretas e outras incorretas e pode ser melhor observado no gráfico da Figura 5.7, onde é constatado o número de acertos e também o número de falsos positivos e negativos, obtidos pela expressão *d=f-g*, se *g=0* e *d=0* então *g=0,2*.

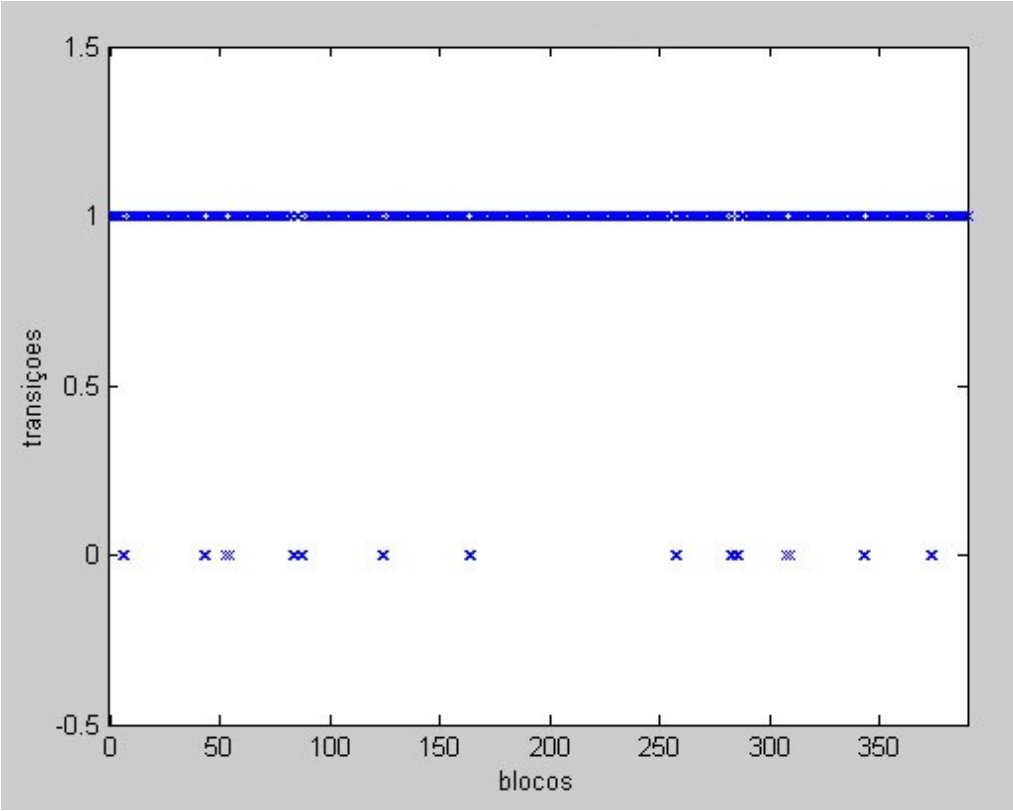


Figura 5.5: Transições manuais, q, de cenas vídeo 1

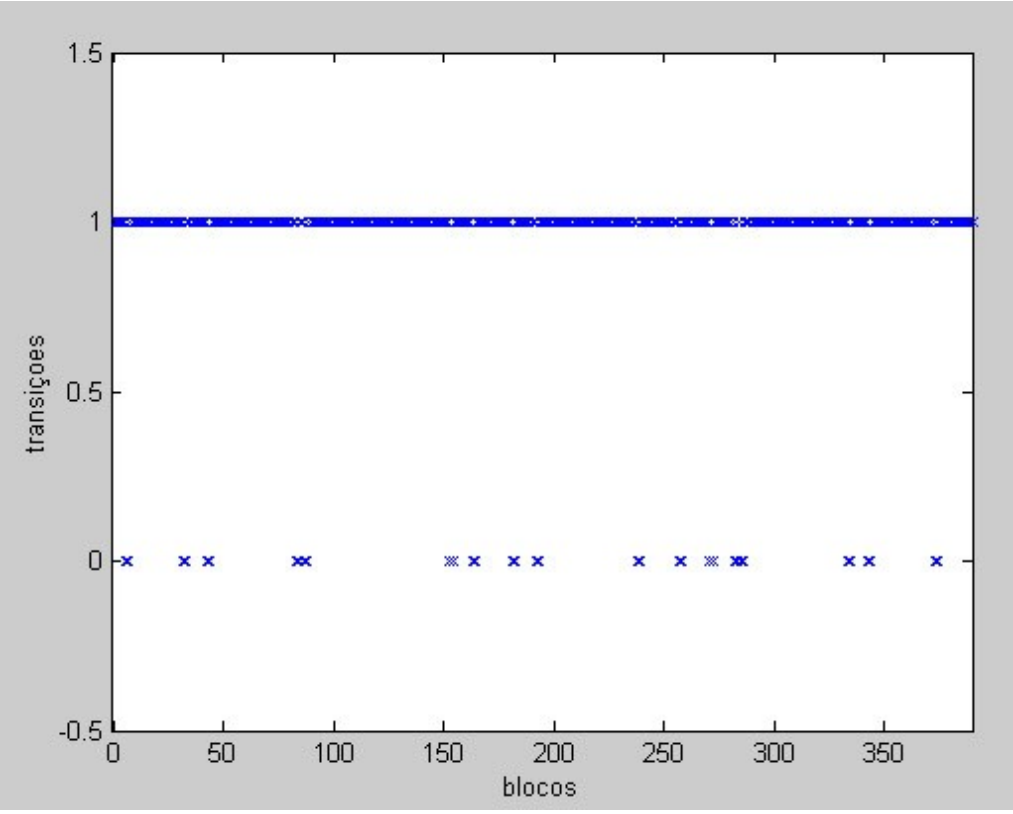


Figura 5.6: Transições automáticas, f, de cenas vídeo 1

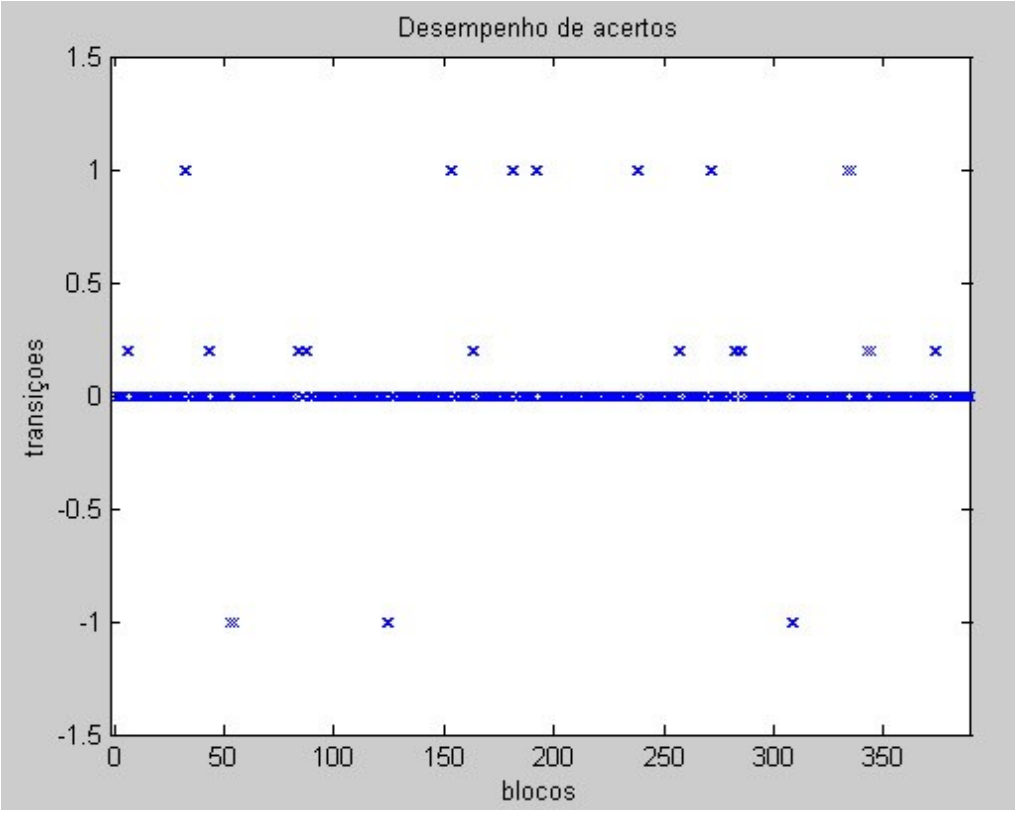


Figura 5.7: Desempenho do vídeo 1

Onde os falsos positivos são representados quando $d(i)=1$, os falsos negativos são representados quando $d(i)=-1$ e os acertos são definidos quando $d(i)=0,2$. Quando $d(i)=0$ significa que não há transição, ou seja estão agrupados em tomadas e cenas.

Dentre as 13 transições que o vídeo 1 possui, o sistema detectou 10 transições corretas, apresentou 7 falsos positivos e 3 falsos negativos. Pode parecer que no primeiro falso positivo há duas transições, isso se deve ao fato de que a transição não representa um corte seco e sim uma transição gradual, onde os efeitos das transições ocorrem em dois blocos consecutivos. Essa

transição foi detectada manualmente e não foi detectada pelo sistema automaticamente.

Analisando o vídeo 1, que se trata de um noticiário, é possível identificar mudanças razoáveis tanto nas características dos quadros, nas tomadas de câmera quanto no conteúdo sonoro, portanto percebe-se um número maior de falsos positivos. A Figura 5.8 ilustra, através de um *waveform* o áudio de uma cena do vídeo 1, onde é possível observar que, embora uma parte do áudio se mantém constante, há variações como o silêncio, por exemplo, dentro da mesma tomada. Quando uma quebra num frame de áudio é feita paralelamente a uma mudança de características visuais dos quadros, o sistema interpreta essa mudança como sendo uma nova cena. Esses resultados foram para um limiar padrão de 0.99998. Esse vídeo foi testado variando p limiar para 0.7 obtendo-se o mesmo resultado.

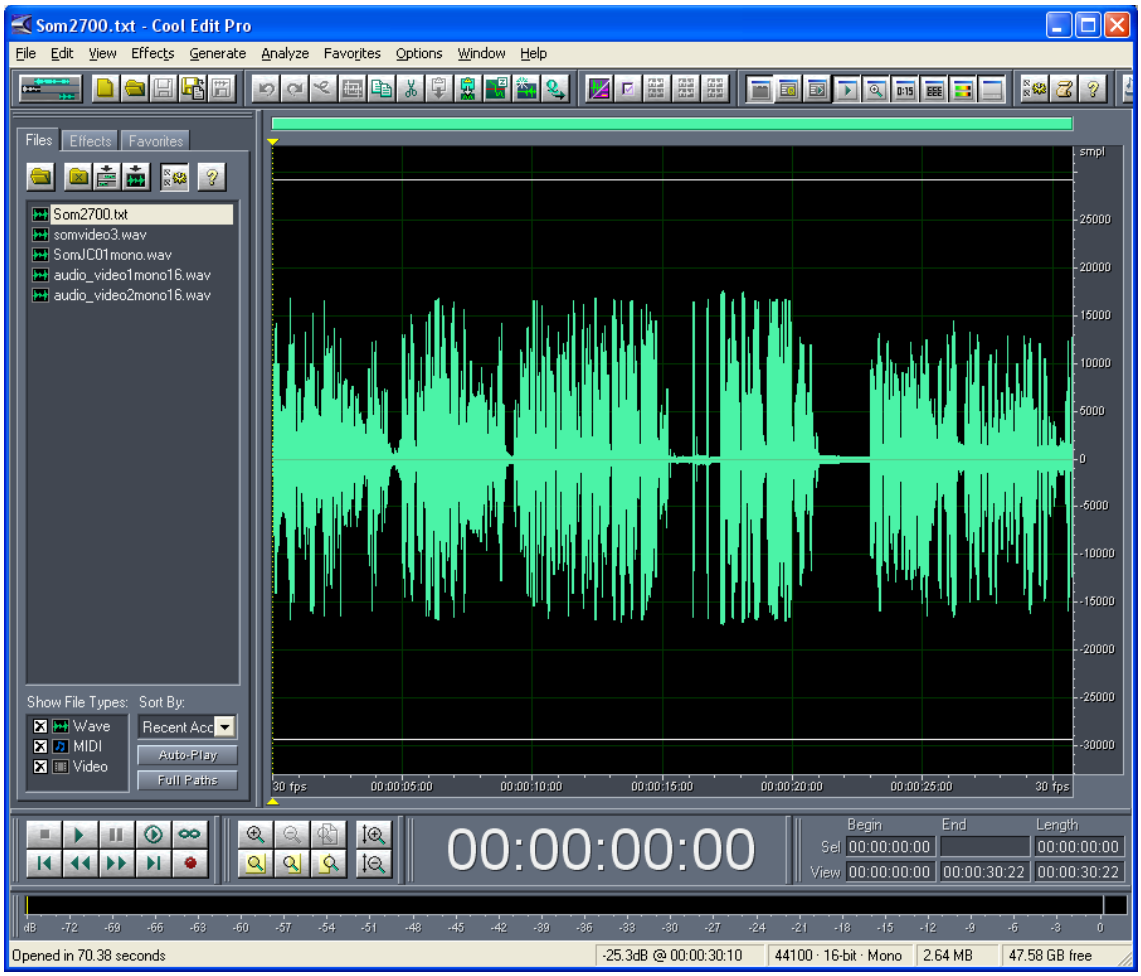


Figura 5.8: Áudio de uma cena do vídeo 1

5.4 Vídeo 2

Duração: 5 minutos e 45 segundos

Cenas: 8

Transições: 7 (definidas manualmente)

Blocos: 327

Quadros: 9.815

Descrição: jornal com diversas reportagens, intercalando comerciais

No gráfico da Figura 5.9 é apresentado o resultado da análise manual, onde $g(i)=0$ representa uma transição de cena, considerando todas as características. E, abaixo, no gráfico da Figura 5.10, é mostrado o resultado da análise automática para detecção de transição de cenas, onde $f(i)=0$ indica que ocorreu transição no bloco i.

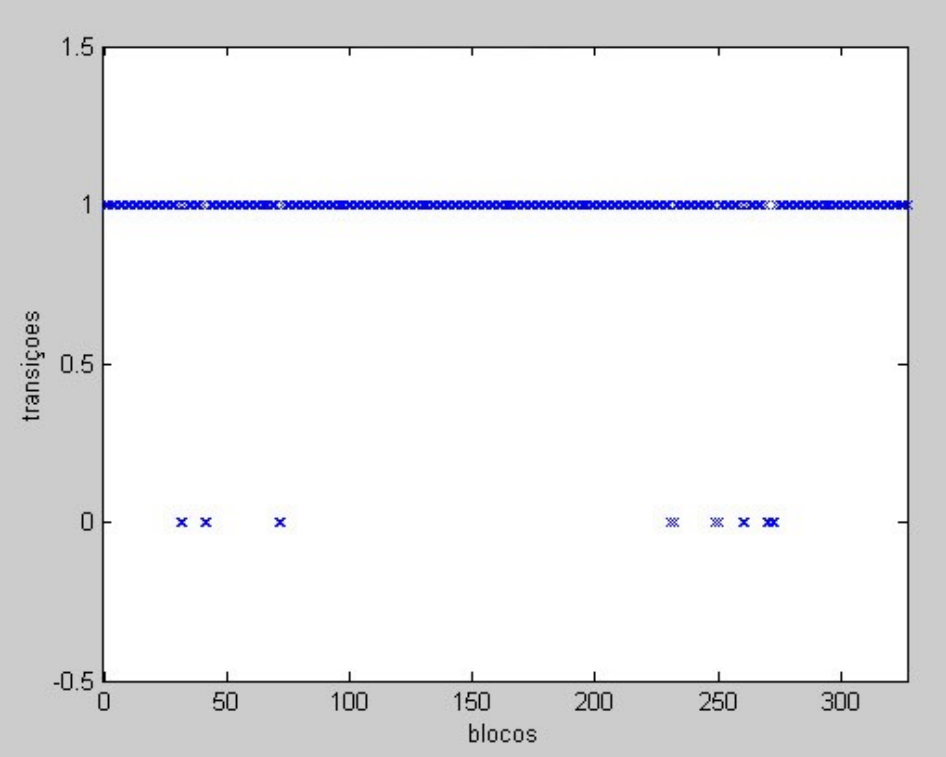


Figura 5.9: Transições manuais, g, de cenas vídeo 2

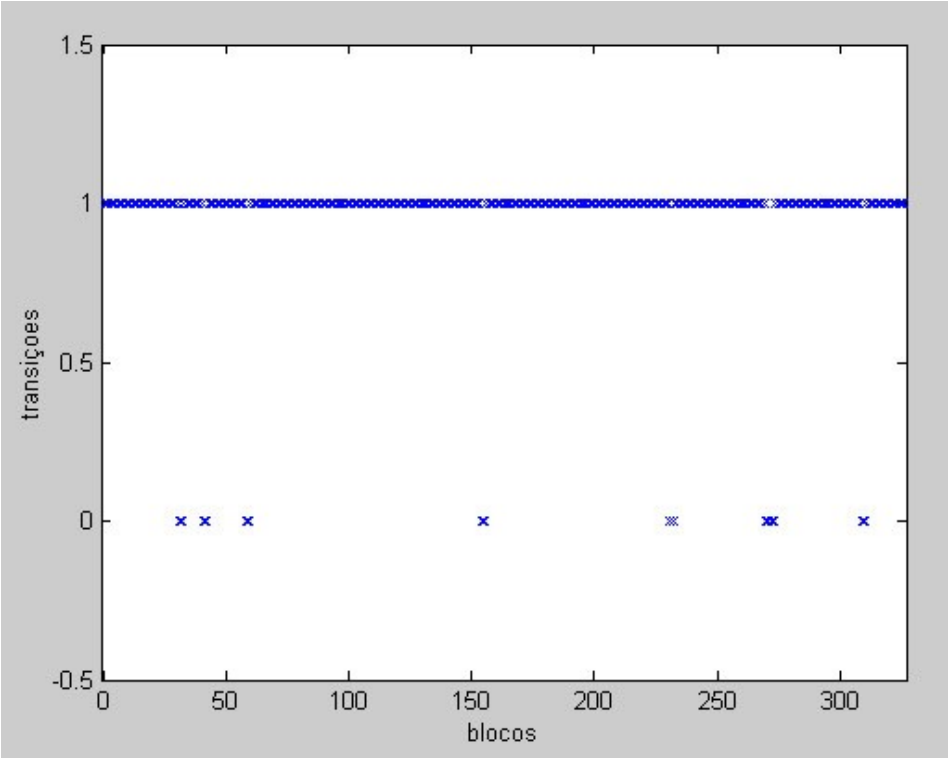


Figura 5.10: Transições automáticas, f, de cenas vídeo 2

A Figura 5.11 mostra o áudio de uma cena do vídeo 2 onde, embora seus quadros possuam características distintas, com transições de corte seco entre as tomadas, como mostra a Figura 5.12, com relação à forma e cor, o áudio permanece seguindo o mesmo padrão.

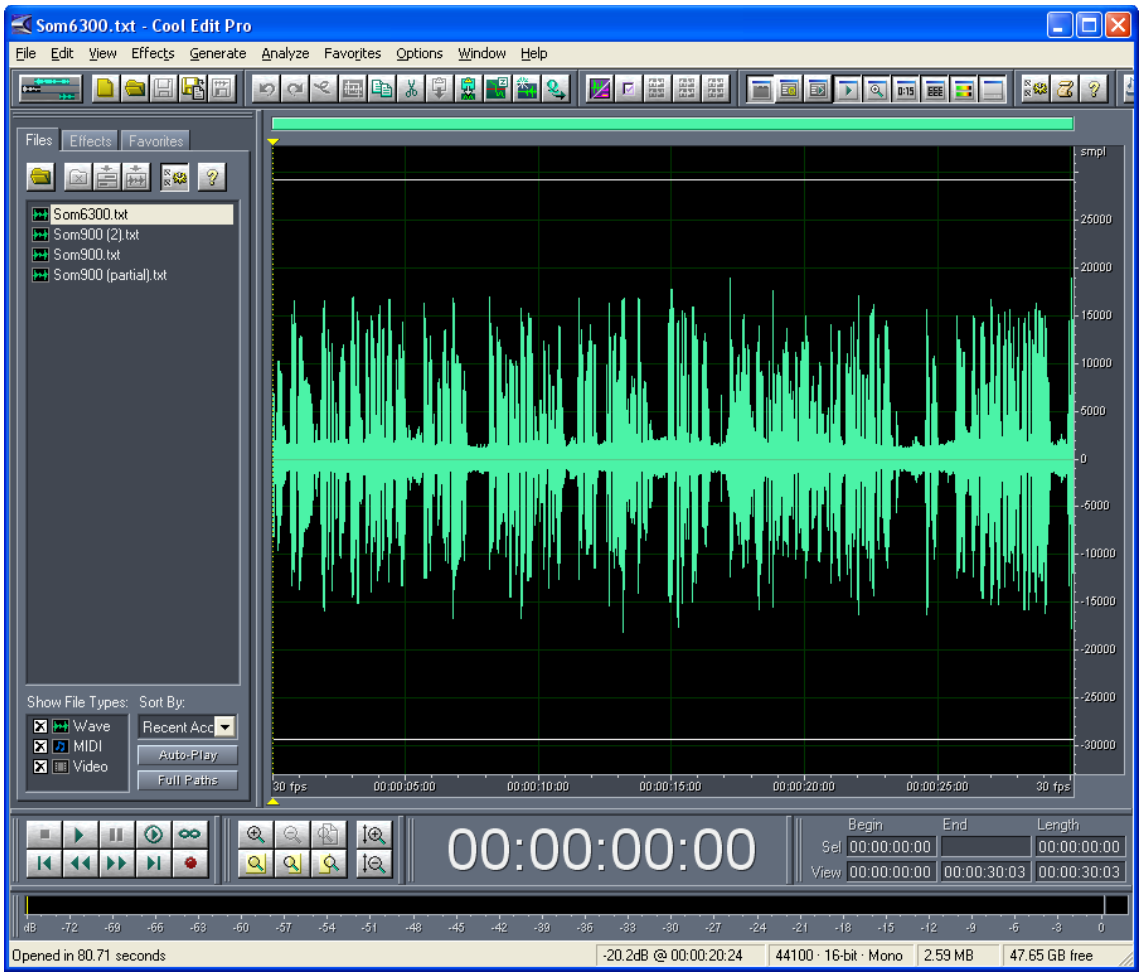


Figura 5.11: Áudio de uma cena do vídeo 2

Figura 5.12: Quadros da mesma cena

O vídeo 2, assim como o vídeo 1, trata-se de um jornal com diversas cenas entre reportagens, entrevistas e comerciais, entretanto, analisando o seu conteúdo manualmente, é possível identificar cenas e tomadas mais definidas e diferenciando-se mais em relação ao vídeo 1. Na Figura 5.13 é possível identificar os acertos, os falsos positivos e os falsos negativos. Os falsos positivos são representados quando $f(i)-g(i)=1$, os falsos negativos são representados quando $f(i)-g(i)=-1$ e os acertos são definidos quando $f(i)-g(i)=0$ e $g(i)$ é uma transição. Neste caso o valor atribuído a $f(i)-g(i)$ é $0,2$, já $f(i)=0$ mostra os quadros que estão agrupados em tomadas e cenas.

Dentre as 7 transições que o vídeo 2 possui, o sistema detectou 5 transições corretas, apresentou 3 falsos positivos e 2 falsos negativos.

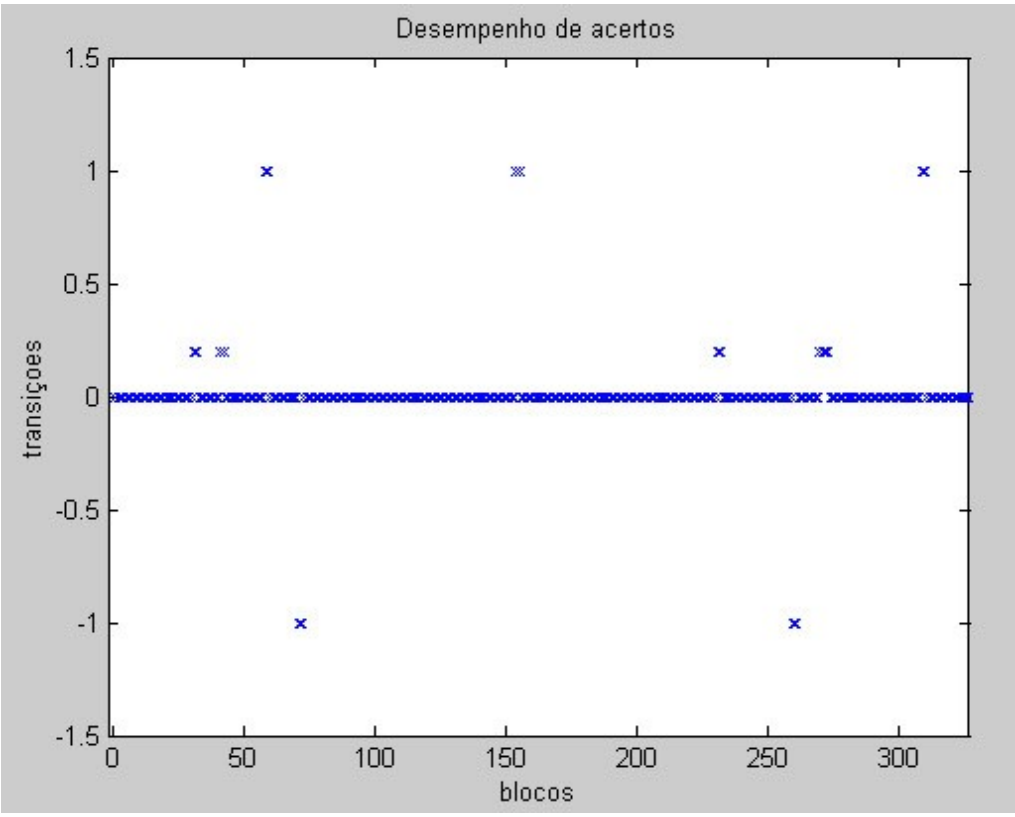


Figura 5.13: Desempenho do vídeo 2

5.5 Vídeo 3

Duração: 2 minutos e 45 segundos

Cenas: 5

Transições: 4 (definidas manualmente)

Blocos: 149

Quadros: 4.496

Descrição: show de banda de rock intercalado com entrevista.

Esse vídeo apresenta um show seguido de entrevista, portanto as características auditivas diferem dos vídeos 1 e 2. Abaixo, na Figura 5.14 é possível visualizar o *waveform* de uma cena do vídeo 3.

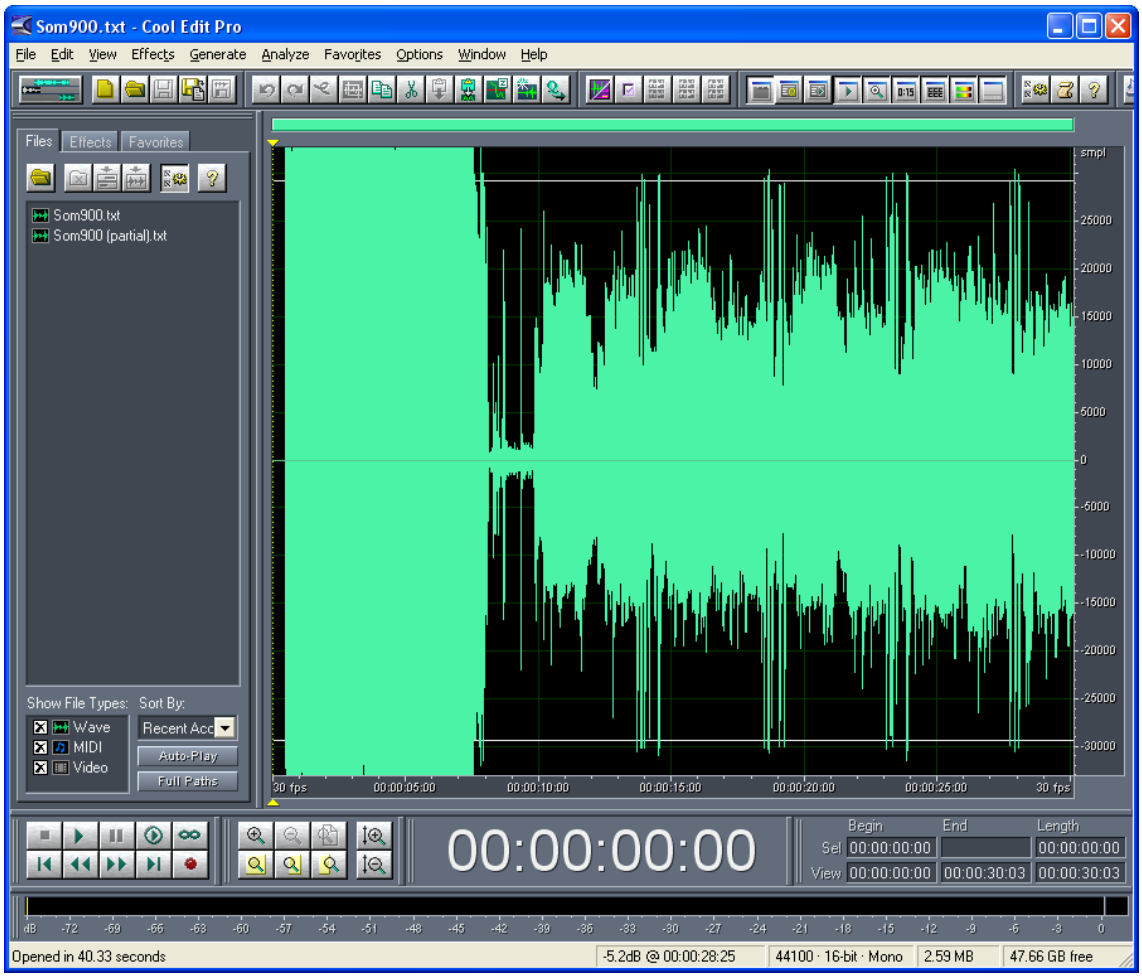


Figura 5.14: Waveform de uma cena do vídeo 3

No gráfico da Figura 5.15 é apresentado o resultado da análise manual, onde $f(i)=0$ representa uma transição de cena. Enquanto no gráfico da Figura 5.16, o mesmo resultado é apresentado, porém se referindo agora à análise automática, gerada pelo sistema.

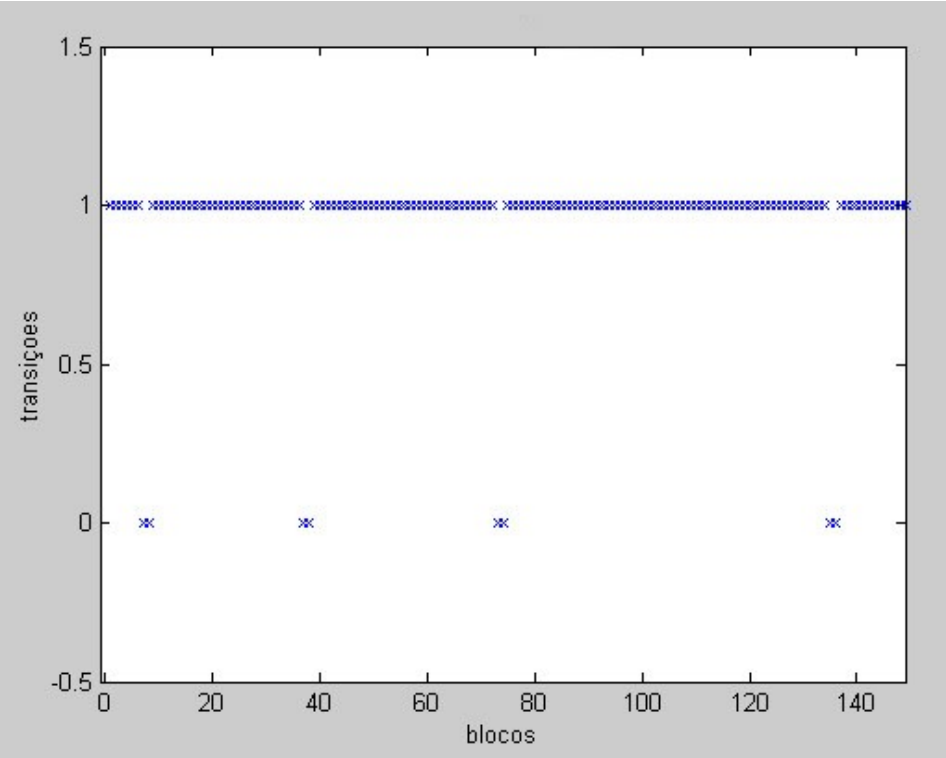


Figura 5.15: Transições manuais de cenas vídeo 3

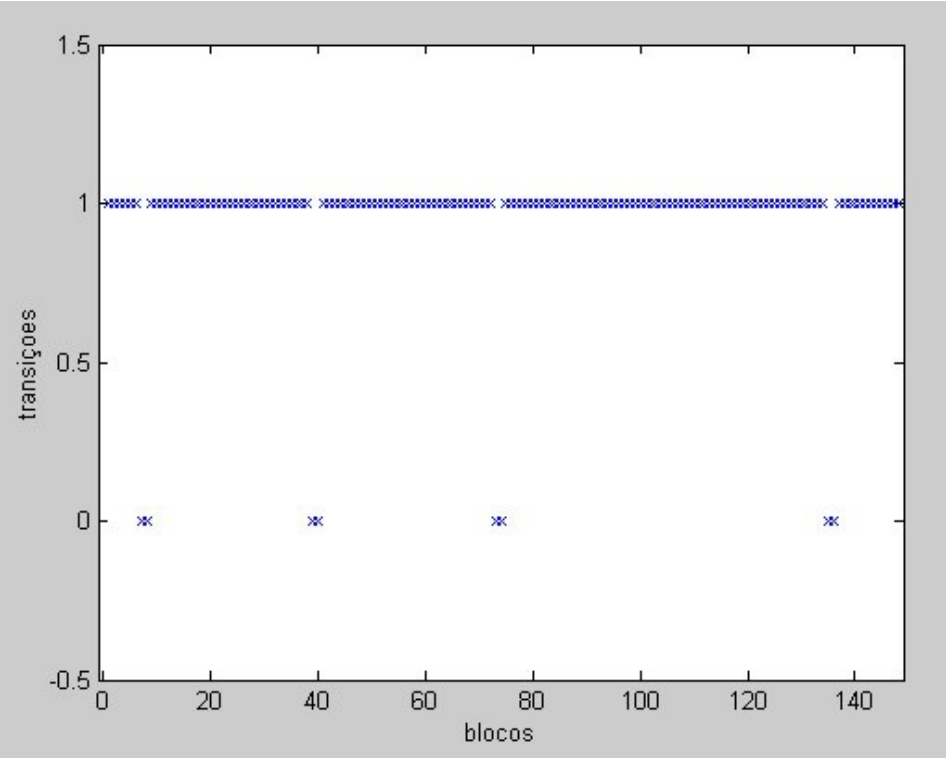


Figura 5.16: Transições automáticas de cenas vídeo 3

No gráfico da Figura 5.17 é possível ver que os acertos absolutos nesse vídeo é alto, pois, entre uma cena e outra, quando há uma quebra no *frame* de áudio, as características das imagens também são alteradas; portanto há alterações paralelas no áudio e nas tomadas, enquanto nas transições entre tomadas, há alterações nas características dos quadros, mas o áudio se mantém constante. O vídeo 3 possui 4 transições de cenas, 3 foram detectadas pelo sistema, entretanto foi incluído um falso positivo e um falso negativo ao redor da real transição.

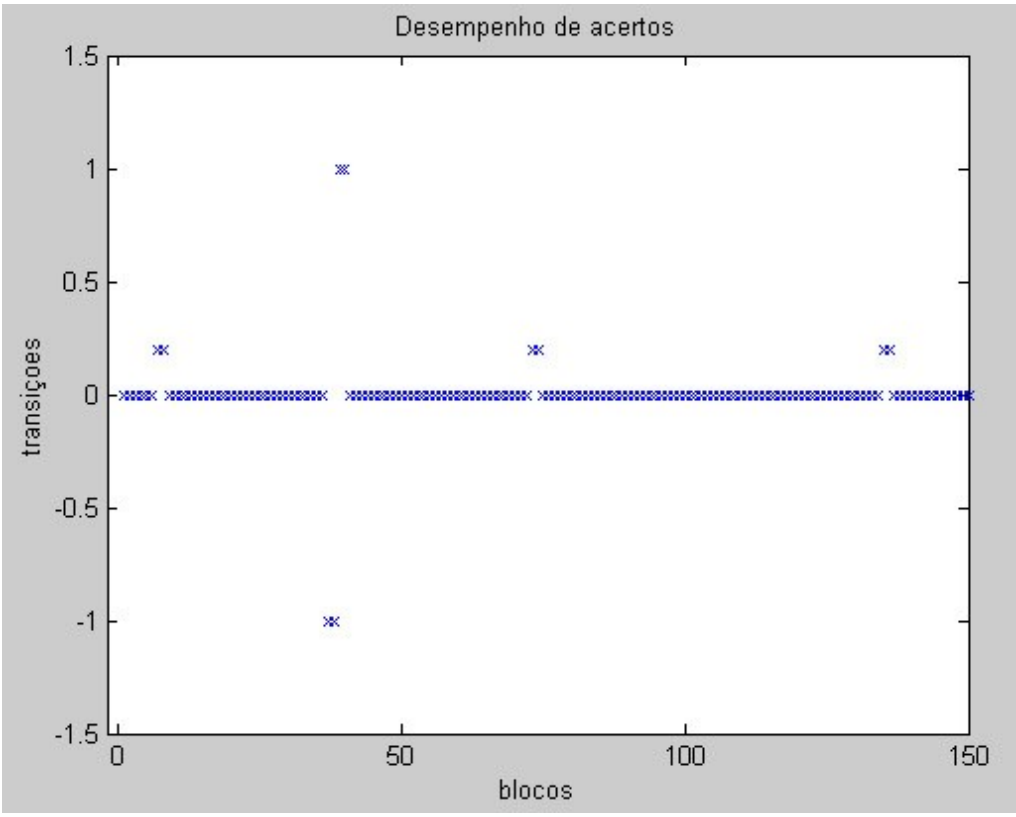


Figura 5.17: Desempenho do vídeo 3

5.6 Vídeo 4

Duração: 5 minutos
Cenas: 27
Transições: 26 (definidas manualmente)
Blocos: 300
Quadros: 9.020
Descrição: desenho animado

No gráfico da Figura 5.18 é apresentado o resultado da análise manual, onde $f(i)=0$ representa uma transição de cena.

No gráfico da Figura 5.19 é possível visualizar as transições de cena automáticas, definidas pelo algoritmo. Quando as Figuras 5.18 e 5.19 são comparadas, são identificadas algumas transições corretas e outras incorretas. Analisando, entretanto o gráfico da Figura 5.20 constata-se que o número de acertos e também o número de falsos positivos e negativos.

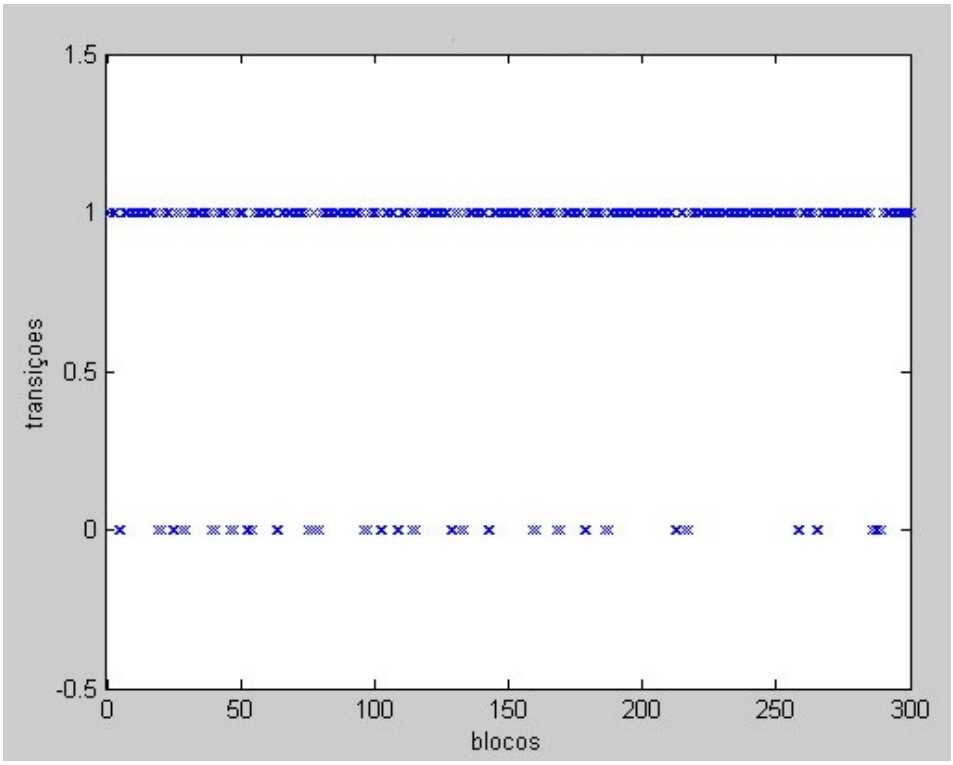


Figura 5.18: Transições manuais de cenas vídeo 4

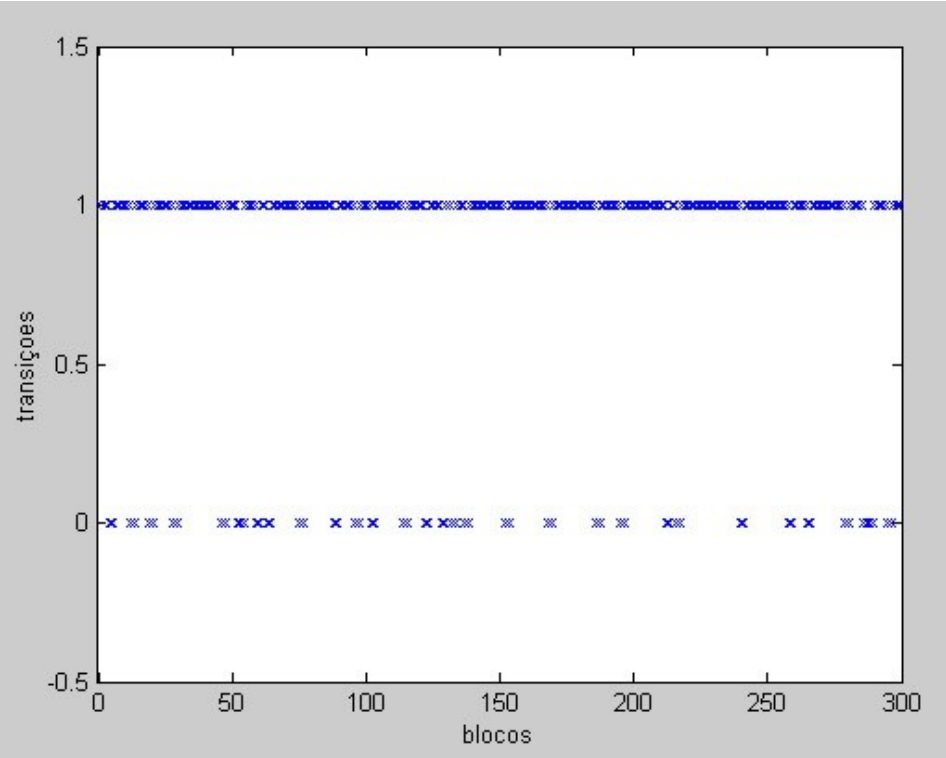


Figura 5.19: Transições automáticas de cenas vídeo 4

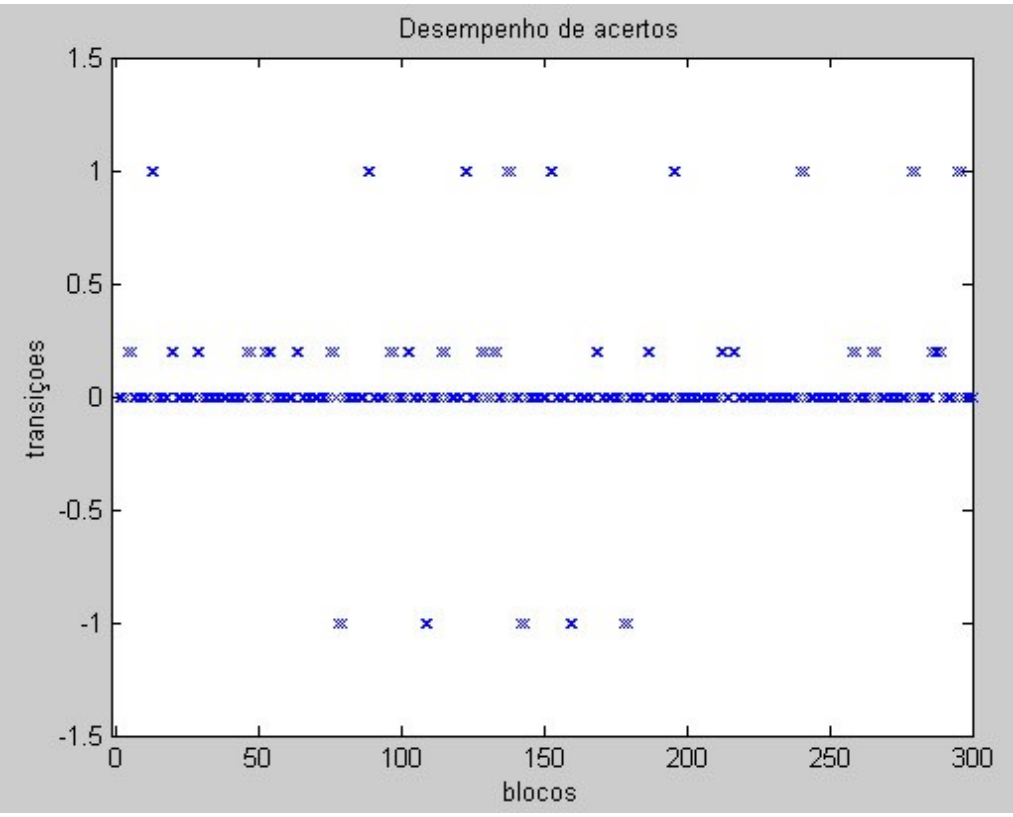
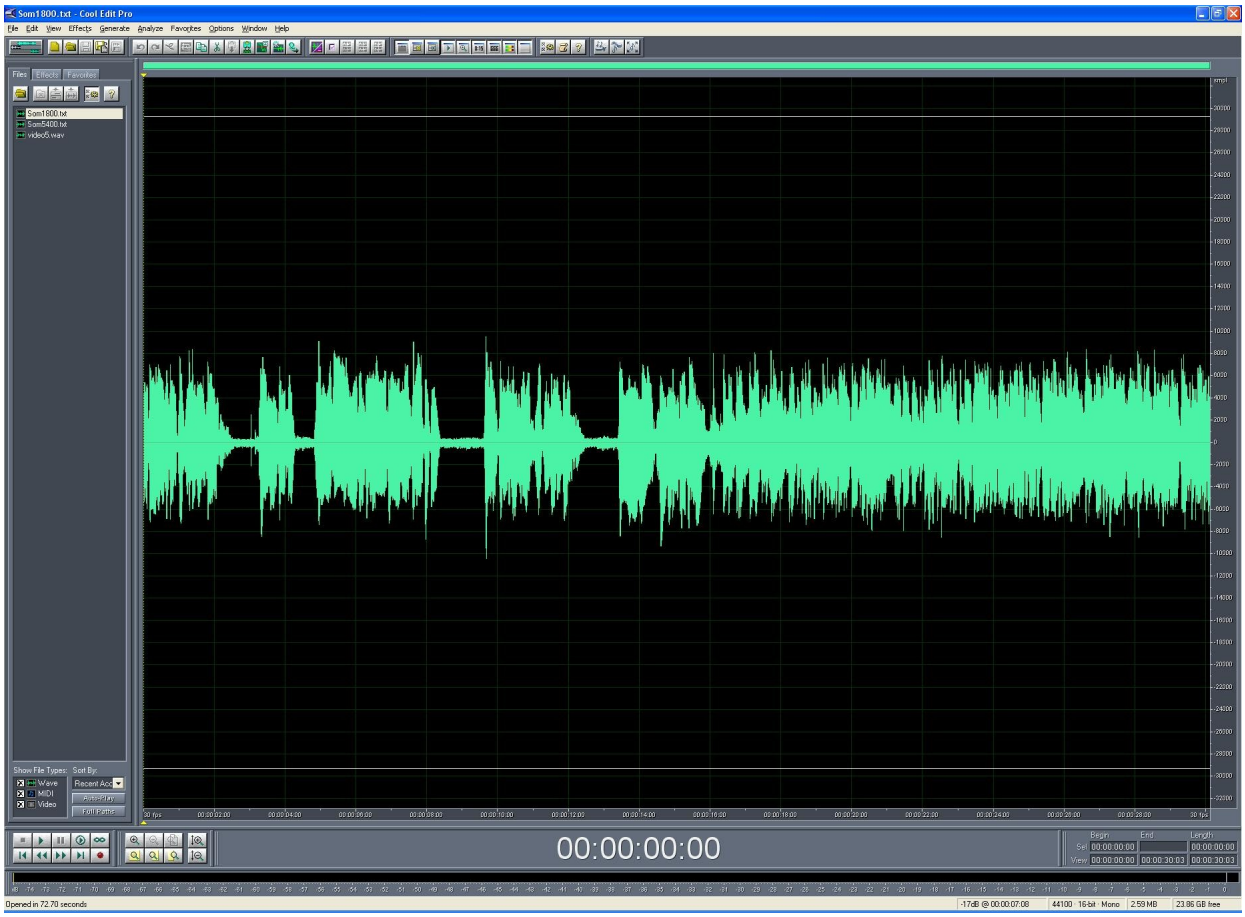


Figura 5.20: Desempenho do vídeo 4

O vídeo 4 é um desenho animado e, embora a detecção automática tenha acertado uma boa parte das transições, também há muitos falsos negativos, ou seja, o sistema detecta transições onde não há. Abaixo, na Figura 5.21 é possível visualizar o *waveform* de uma cena do vídeo 4 que apresenta o áudio mais ou menos constante, ao longo da cena.

Dentre as 26 transições que o vídeo 4 possui, o sistema detectou 20 transições corretas, apresentou 10 falsos positivos e 7 falsos negativos.



.

Figura 5.21: Waveform de uma cena do vídeo 4

5.7 Vídeo 5

Duração: 5 minutos

Cenas: 7

Transições: 6 (definidas manualmente)

Blocos: 300

Quadros: 9.003

Descrição: documentário brasileiro

No gráfico da Figura 5.22 é apresentado o resultado da análise manual, onde $f(i)=0$ representa uma transição de cena.

No gráfico da Figura 5.23 é possível visualizar as transições de cena

automáticas, definidas pelo algoritmo. Quando as Figuras 5.22 e 5.23 são comparadas, são identificadas algumas transições corretas e outras incorretas. Analisando, entretanto o gráfico da Figura 5.24 é constatado o número de acertos e também o número de falsos positivos e negativos.

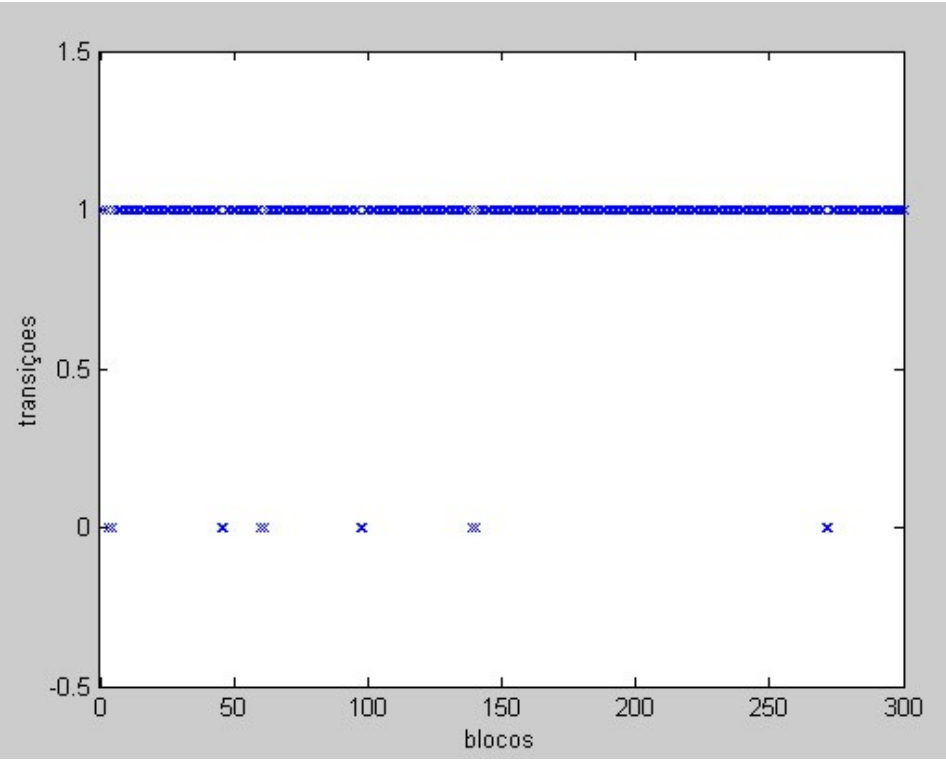


Figura 5.22: Transições manuais de cenas vídeo 5

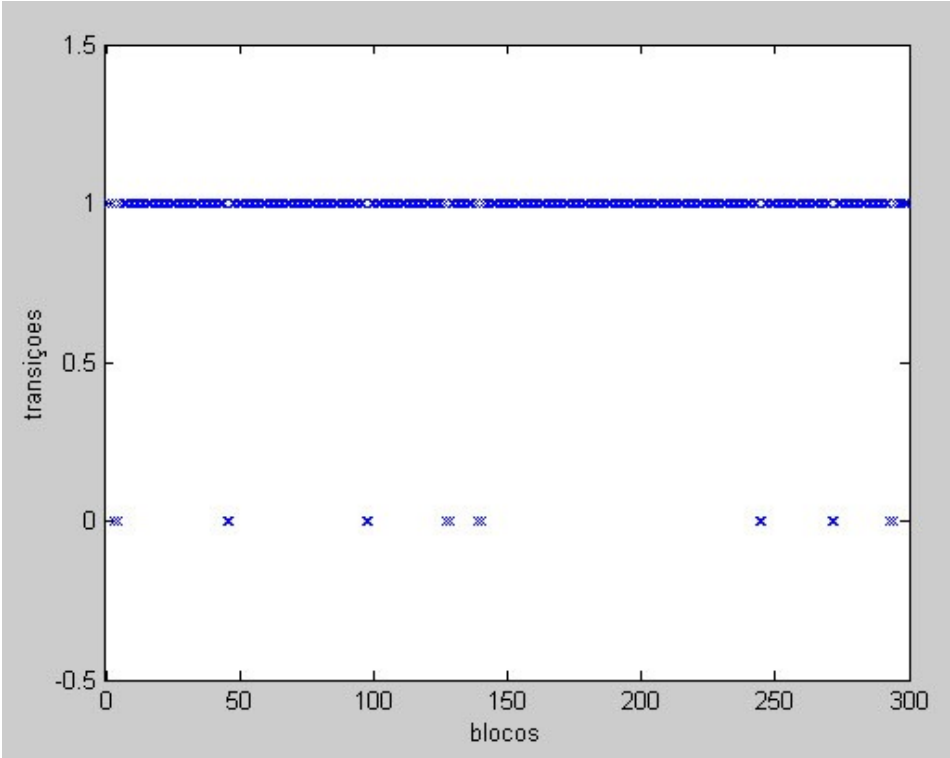


Figura 5.23: Transições automáticas de cenas vídeo 5

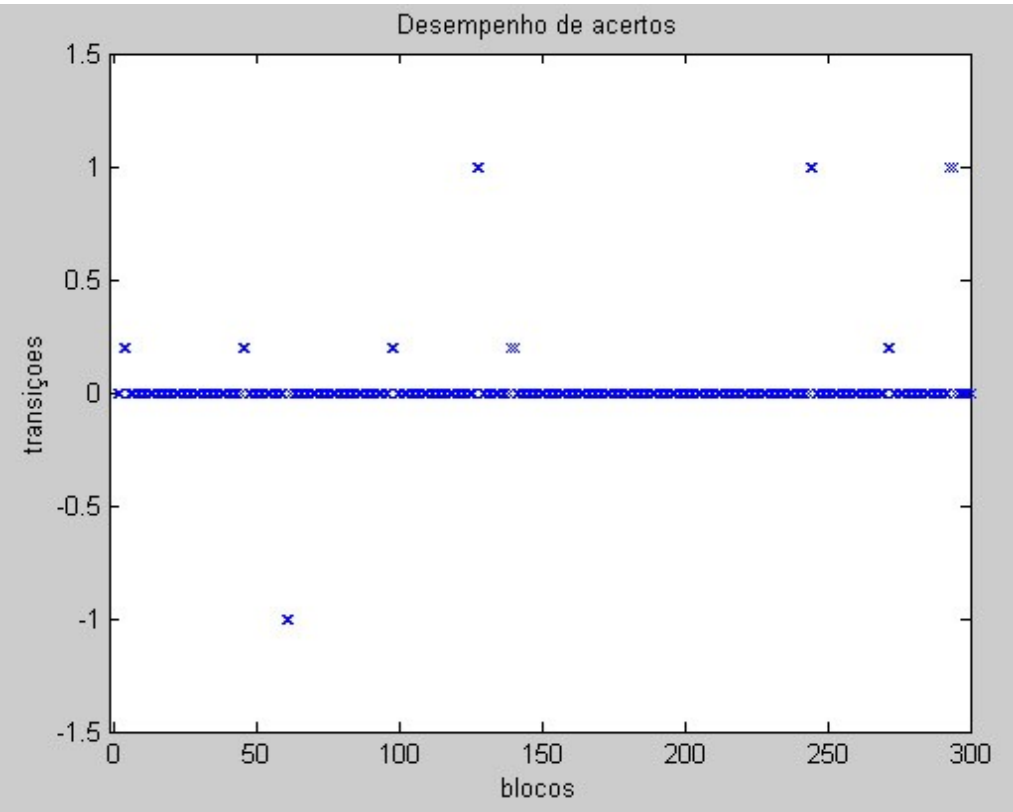


Figura 5.24: Desempenho do vídeo 5

O vídeo 5 mostra um filme nacional que apresenta cenas e tomadas bem definidas. O desempenho do algoritmo é considerado bom, visto que das 6 transições existem, 5 são detectadas. Há apenas 1 falso negativo e 3 falsos positivos.

Abaixo, na Figura 5.25 é possível visualizar o *waveform* de uma cena do vídeo 5. Este vídeo também foi testado usando um limiar de 0.7, onde obteve-se um resultado melhor.

Figura 5.25: Waveform de uma cena do vídeo 5

5.8 Análises

De acordo com o que foi visto no presente capítulo, a utilização do algoritmo provou ser uma boa ferramenta para detectar as transições de cena de um vídeo. No exemplo do vídeo 3, foi obtido um maior número de acertos em valores absolutos, se comparado aos demais, mas a maior porcentagem de acertos ocorrem no vídeo 5. A tabela 3.8 faz uma comparação desses percentuais:

Tabela 3.8: Tabela de percentuais nos experimentos

Vídeos	% Acertos	Falsos Positivos	Falsos Negativos
1	76,92	53,85	23,08
2	71,43	42,86	28,57
3	75	25	25
4	76,92	38,46	26,92
5	83,33	50	16,67

Nesse trabalho, o áudio não foi classificado, por exemplo, como fala, silêncio, uma pessoa falando, várias pessoas falando, ruído, música etc., O objetivo principal foi comparar os segmentos de áudio ao longo do vídeo sem a necessidade de classificá-lo e sim, compará-lo aos segmentos subseqüentes. Os resultados não indicam, na verdade, como as características se correlacionam.

Supõe-se que o áudio nem sempre indicaria a detecção de mudança de cena. No experimento 4, por exemplo, onde foi analisado um desenho animado, o áudio se manteve constante em praticamente todas as cenas, isso indica que, possivelmente, em situações como essa, no caso dos desenhos animados, um peso menor pode ser atribuído às características auditivas aumentando, com isso, o peso atribuído às características dos quadros.

Nos experimentos, foi utilizado um limiar de 0.999998, entretanto, quando foi alterado para 0.7 foi possível notar que, em alguns vídeos como os noticiários (experimentos 1 e 2) o resultado final não foi alterado, entretanto no filme, por exemplo, o número de acertos foi melhor, o algoritmo detectou 6

transições, quando, na verdade, haviam 7 transições e apenas 2 falsos positivos.

6. Conclusões e Trabalhos Futuros

Neste trabalho foi apresentado um modelo que combina diversas características para indexação de vídeos combinando cor, forma e som.

A detecção de cena automática é uma tarefa desafiadora. Usando apenas a informação de áudio ou de cor e forma, separadamente, pode-se não chegar a resultados satisfatórios, entretanto combinando informação de forma, cor e áudio eficientemente, pode melhorar muito o resultado das transições, isso porque as fronteiras ao redor das tomadas dos quadros nem sempre representam as fronteiras de uma única característica.

Para a detecção de transições de cenas, não há necessidade de focar qual é o conteúdo do áudio (música, fala), ou da imagem, ao invés disso, o foco é a diferença entre as tomadas e cenas de vídeo.

A utilização de redes bayesianas contribui com a etapa de segmentação de cenas, podendo ser facilmente incluídas novas abordagens tais como textura e no conteúdo visual e quaisquer outras características sonoras, sem que isso envolva grandes alterações. Além disso, também é possível atribuir pesos diferentes às diversas características, facilitando experimentos e testes futuros.

Três vídeos foram realizados com dois limiares alterando o um dos valores para 0,7. Esse resultado se mostrou mais eficiente para a análise das transições de cena do vídeo 5, que se refere a um filme. Nesse sentido, uma sugestão para trabalhos futuros é fazer um estudo dos limiares que se adequam mais a cada tipo de vídeo, por exemplo, filmes, desenhos, noticiários, jogos esportivos. Nesse sentido é sugerido um estudo para identificar o melhor limiar para cada tipo de vídeo.

Além disso, foi utilizado o tamanho de bloco igual a 5. Novas propostas

de pesquisa podem incluir, também, testes que alteram o tamanho dos blocos para a obtenção de possíveis melhoras nos resultados.

A técnica implementada apresenta diversas limitações principalmente quanto ao número de características utilizadas, por exemplo, foram utilizadas apenas uma quantização de cor e um espaço de cor. Obviamente, poderiam ser utilizadas outras quantizações no mesmo espaço HSV. Esse espaço poderia ter sido trocado por outros, por exemplo RGB, LUV etc.

O vetor de característica de formas se baseou apenas no cálculo das tangentes dos pontos de contorno. Essa quantização se baseou num histograma de 10 posições. Outras quantificações do vetor tangente poderiam ter sido usadas mas, principalmente, poderiam ter sido usadas outras técnicas como uso de momentos invariantes, descritores de Fourier e diversas outras conhecidas na literatura para descrição de formas dos quadros.

O mesmo ocorre com as características de som onde diversas outras características poderiam ter sido usadas como a descrição do som no domínio da freqüência.

Todas essas mudanças poderiam tornar o trabalho mais completo e mais complexo no sentido de promover pesquisas em processamento e análise de imagens podendo, também, ser utilizado na linha de pesquisa em trabalhos futuros.

O problema da classificação contínua de dados em um sistema, como visto em [LI01], está principalmente na escolha das características a serem utilizadas. Haja visto que as técnicas sonoras melhoram os resultados, o presente trabalho sugere que seja feito um estudo mais aprofundado nas diversas características de extração sonoras existentes para que, a partir desse resultado, seja possível a obtenção de resultados melhores e mais precisos em vídeos segmentados.

A partir dessa abordagem, muitos outros trabalhos poderão ser desenvolvidos também, na comparação das redes bayesianas com outros

modelos como, por exemplo, redes neurais.

Referências Bibliográficas

[BALAKRISHNAN97] V. K. Balakrishnan. Theory and Problems of Graph Theory. Editora McGraw-Hill, 1997.

[BRICE70] C. Brice and C. Fennema. Scene analysis using regions. Artificial Intelligence, pp. 205–226. 1970.

[CASTRO98] E. M. M. M. de Castro. Recuperação de Imagens em Banco de Dados por Semelhança de Cor, Computação Aplicada e Automação – CAA-UFF, dissertação de mestrado. 1998.

[COELHO04] T. Coelho, P. Calado, L. Souza and B. Ribeiro Nt. Image retrieval using multiple evidence ranking. IEEE Transactions on Knowledge and Data Engineering, pp. 408 – 417, 2004.

[CONCI99] A. Conci and E. M. M. M. Castro. Color image retrieval systems: a comparison of appraches, Anais do V Simpósio Brasileiro de Multimídia e Hipermídia – SBMIDIA'99 – Sociedade Brasileira de Computação, organizado pelo Instituto de Informática da UFG, Goiânia, pp. 141-159, 1999.

[CONCI00] A. Conci and E. M. M. M. Castro. Image retrieval system in databases: the problem of color similarity. Proceedings of 9th International Conference on Geometry and Graphics, Edited by Johann Pretorius, Published by Rand Afrikaans University Johannesburg, pp. 174-178. South Africa, 2000.

[CONCI01] A. Conci and E. M. M. M. Castro. Image mining by color content, proceedings of ACM International Conference on Software Engineering & Knowledge Engineering. Buenos Aires, 2001.

[CONCI02] A. Conci and E. M. M. M. Castro. Image mining by content. Journal of Experts Systems with Applications. Elsevier Science UK, volume 4, pp. 377-

383. 2002.

[EAKINS02] J. P. Eakins. Towards intelligent image retrieval. Pattern Recognition, pp. 3–14, 2002.

[FISHER95] S. Fisher, R. Lienhart and W. Effelsburg. Automatic recognition of film genera. Proceedings of Multimedia-95. ACM Press. San Francisco. pp. 295-305. Novembro 1995.

[FONSECA96] J. S. Fonseca and G. A. Martins. Curso de Estatística**.**. Editora Atlas 6ª edição, cidade São Paulo, 1996.

[FOOTE99] J. Foote. Visualizing Music and Audio using Self Similarity. In Proceedings of ACM on Multimedia. 1999.

[GOUYON00] F. Gouyon, F. Pachet and O. Delerue. On the use of zero-crossing rate for an application of classification of percussive sounds. Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Dezembro 2000.

[HERMES95] T. Hermes, C. Klauck, J. KreyB and J. Zhang. Image retrieval for information systems. Store and Retrieval for Image and Video Databases III, Proceedings of SPIE, volume 2420, pp. 394–405, Fevereiro 1995.

[HOROWITZ97] C. E. Horowitz, S. Sahini and S. Rajasekaran. Computer Algorithms. Computer Science Press, 1997.

[HRUSCHKA97] E. R. J. Hruschka. Propagação de Evidências em Redes Bayesianas: Diagnóstico sobre Doenças Pulmonares. Dissertação de mestrado, Universidade de Brasília, Departamento de Ciência da Computação, Brasília, 1997.

[HU62] M. Hu. Visual pattern recognition by moment invariants. IEEE Transaction on Information Theory, IT-8. 1962.

[JAVED00] O. Javed, S. Khan, Z. Rasheed and M. Shah. Framework for Segmentation of Interview Videos. Computer Vision Lab. University of Central Florida. Florida, 2000.

[JENSEN01] F.V.Jensen. Bayesian networks and decision graphs. Statistics for Engineering and Information Science. Springer, 2001.

[JIANG00] H. Jiang, T. Lin and H.J. Zhang. Video segmentation with the assistance of audio content analysis. *IEEE International Conference on Multimedia and Expo (ICME'00),* pp. 1507-1510, 2000.

[KORPI03] J. Korpi-Anttila. Automatic colour enhancement and scene change detection of digital video. Graphics Arts in Finland. Finland. 2003.

[LEE96] T.S. Lee, Image representation using 2D Gabor wavelets, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.18, Outubro 1996.

[LI01] D. Li. I. K. Sethi, N. Dimitrova and T. McGee. Classification of general audio data for content-based retrieval. Pattern Recognition Letters, pp. 533-544, Abril 2001.

[LIENHART97] R. Lienhart, S. Pfeiffer and W. Effelsberg. Video abstracting. Communications of ACM pp. 54-62. 1997.

[LIPSCHUTZ68] S. Lipschutz. Theory and problems of probability, Editora McGraw Hill, Cidade: New York, 1968.

[LIPSON97] P. Lipson, E. Grimson and P. Sinha. Configuration-based scene classification and image indexing. Proceedings of IEEE Conference on Computer Visiob and Pattern Recognition (CVPR'97), pp. 1007–1013, Puerto Rico, 1997.

[LIU98] Z. Liu, Y. Wang and T. Chen. Audio feature extraction and analysis for scene segmentation and classification. Journal of VLSI Signal Processing

Systems. Junho 1998.

[LU02] L. Lu, H. J. Zhang, and H. Jiang. Content analysis for audio classification and segmentation. IEEE Transaction on Speech and Audio Processing, pp. 504-516, Outubro 2002.

[MATHIAS98] E. Mathias and A. Conci. Comparing the influence of color spaces and metrics in content-based image retrieval, SIBGRAPI'98 – Proceedings of the International Symposium on Computer Graphics, Image Processing and Vision – published by IEEE Computer Sciente – pp. 371 – 378. Rio de Janeiro, 1998.

[MURAMOTO00] T. Muramoto and M. Sugiyama. Visual and audio segmentation for video streams. *IEEE International Conference on Multimedia and Expo (ICME'00),* pp 1547-1550, 2000.

[NAGASAKA92] A. Nagasaka and Y. Tanaka .Automatic video indexing and full-video search for object appearances.*IFIP. Proceedings, Visual Database Systems*, vol. 2, pp. 113-127. Elsevier Science Publishers B.V. North-Holland, 1992.

[NAIM04] P. Naim, P. Wuillemim, P. Leray, O. Pourret, e A. Becker. Réseaux Bayésiens. Eyrolles, Paris, 2004.

[OLIVA99] A. Oliva, A. B. Torralba, A. Guerin-Dugue and J. Herault. Global semantic classification of scenes using power spectrum templates. In The Challenge of Image Retrieval. CIR-99, Newcastle upon Tyne, UK, Fevereiro 1999.

[PARASKEVAS03] I. Paraskevas and E. Chilton. Áudio classification using acoustic images for retrieval from multimedia databases. EURASIP Conference focused on Video / Image Processing and Multimedia Communications. Croatia. Julho 2003.

[PEARL88] J. Pearl. Probabilistic Reasoning in Intelligent Systems. Morgan

Kaufmman, Los Autos, California, 1988.

[PRATT91] W. Pratt. Digital image processing, volume 2. Wiley-Interscience. 1991.

[RASHEED02] Z. Rasheed and M. Shah A Graph Theoretic Approach for Scene Detection in Produced Videos. Computer Vision Lab. University of Central Florida. Florida, 2002.

[REN02] W. Ren and S. Singh. Video transition: modelling and prediction. IEEE Trasaction on Speech and Audio Processing, pp. 536-549, Outubro 2002.

[RIBEIRO96] B. Ribeiro Nt and Richard R. Muntz. A belief network model for IR. In ACM Conference on Research and Development in Information Retrieval – SIGIR'96, pp. 253–260, 1996.

[RIBEIRO98] B. Ribeiro Nt, G. G. Guimarães, and I. R. Silva. Clipping: A technique for improving retrieval performance. In In XVIII International Conference of Chilean Society of Computer Science, pp. 140–148, 1998.

[RIBEIRO00] B. Ribeiro Nt, Ilmério Silva and Richard Muntz. Bayesian network models for IR. In Fabio Crestani and Gabriella Pasi, editors, Soft Computing in Information Retrieval Techniques and Applications, pp. 259–291, Editora Springer Verlag, 2000.

[RODRIGUES03] P. S. Rodrigues. Um Modelo Bayesiano Combinando Análise Semântica Latente e Atributos Espaciais para Recuperação de Informação Visual. Tese de doutorado. UFMG Belo Horizonte, Maio 2003.

[RODRIGUES05] P. S. Rodrigues, A. A. Araújo and G. A. Giraldi. Using Tsallis entropy a bayesian network for CBIR, Proceeding of International Conference on Image Processing (ICIP'05). Genova, Itália, Setembro 2005.

[RUI99] Y. Rui, T. Huang and S. Chang. Image retrieval: current techniques, promising directions, and open issue. *Journal of Visual Comm. and Image*

Representation, pp. 39–62. 1999.

[RUSSELL95] S. Russell and P. Norving. Artificial intelligence: a modern approach. Prentice Hall, 1995.

[SCHEIRER97] E. Scheirer and M. Slaney. Construction and evaluation of a robust multi feature speech/music discriminator. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings, pp. 1331-1334, Abril 1997.

[SHEN96] K. Shen and E. Delp. A fast algorithm for video parsing using MPEG compressed sequences. Proceedings of International Conference on Image Processing (ICIP'96). Lausanne, 1996.

[SILVA00] I. Silva, B. Ribeiro-Neto, P. P Calado, E. S. Moura, and N. Ziviani. Link-based and content-based evidential information retrieval in a belief network model. Proceedings of the 23rd Annual ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 96–103, Julho 2000.

[SMITH96] J. R. Smith and S. F. Chang. Visualseek: a fully automated content-based query system. ACM Multimedia 96, pp. 87–98, Boston, M.A., 1996.

[SMITH97] J. R. Smith and S. F. Chang. A general framework for integrated spatial and feature image search. In*Workshop on Multimedia Signal Processing*, pp. 301–306. Boston, M.A., 1997.

[SPINA96] M. Spina and V. W. Zue. Automatic transcription of general audio data: preliminary analyses. Proceedings International Conference of Spoken Language Processing, pp. 594-597. Philadelphia. Outubro 1996.

[SUNDARAM00] H. Sundaram and S. F. Chang. Video Scene Segmentation using video and audio features. I*EEE International Conference on Multimedia and Expo (ICME'00),* pp. 1145-1148, 2000.

[SZUMMER98] M. Szummer and R. Picard. Indoor-outdoor image

classification. In IEEE International Workshop on Content-based Access of Image and Video Databases (CAIVD'98), pp. 42–51, Bombay, India, 1998.

[THOMPSON00] S. Thompson, D. S. Daly, E. M. Perry and K. K. Anderson. Local linear statistical modeling as a framework for change detection. Laboratory-Directed Research and Development program at Pacific Northwest National Laboratory, 2000.

[TONG01] C. C. Tong and J. Kuo. Audio Content Analysis for Online Audiovisual Data Segmentation and Classification. IEEE Trasaction on Speech and Audio Processing, volume 9, pp. 441-454, Maio 2001.

[TURTLE91] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. ACM Transactions on Information Systems, pp. 187–222, 1991.

[TZANETAKIS02] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, pp. 293-302, Julho 2002.

[VAILAYA98] A. Vailaya, A. Jain and H. J. Zhang. On image classification: city images vs landscapes. Pattern Recognition, pp. 1921—1936, 1998.

[VASCONCELOS97] N. Vasconcelos and A. Lippman. Bayesian video shot segmentation. MIT Media Laboratory, 1997.

[YATES99] R. B. Yates and B. Ribeiro Nt. Moderm Information Retrieval. Editora Addison Wesley, Cidade Harlow, 1999.

[YOO02] H. W. Yoo, D. S. Jang, S. H. Jung, J. H. Park and K. S. Song. Visual information retrieval system via content-based approach. Pattern Recognition, pp. 749–769, 2002.

[YOSHITAKA01] A. Yoshitaka and M. Miyake. Scene Detection by Audio-Visual features. *IEEE International Conference on Mulatimedia and Expo*

(*ICME'01*), pp. 49-52, 2001.

[ZABIH99] R. Zabih, J. Miller and K. Mai .Feature-based algorithm for detecting and classifying production effects. *ACM Journal on Multimedia Systems*, vol. 7, pp. 119-128, 1999.

[ZHANG93] H.J. Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, vol. 1, pp. 10-28, 1993.

[ZHANG99] T. Zhang and C. C. J. Kuo. Video content parsing based on combined audio and visual information. SPIE, vol. IV, pp. 78-89,1999.

[ZHANG00] H. J. Zhang, T. Lin, and H. Jiang. Video segmentation with the assistance of audio content analysis. IEEE Trasaction on Speech and Audio Processing, pp. 536-549, Outubro 2000.

[ZHOU01] X. S. Zhou and T. S. Huang. Edge-based structural features for content-based image retrieval. Patterns Recognition Letters, pp. 457–468, 2001.

[ZHU03] Y. Zhu and D. Zhou. Scene change detection based on audio and video content analysis. IEEE International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 03), pp. 695 -702, 2003.

Anexos

O formato .AVI

Embora existam vários formatos de arquivos de vídeo digital, como por exemplo .ASF, .MOV, .MPEG, .RAM, a escolha pelo formato .AVI (*Audio Video Interleave*) para captura, se deve a alguns fatores como por exemplo, sua larga utilização e menor compressão dentre os demais.

Devido ao fato do formato .AVI não ser comprimido por um ***CODEC*** (*Compression/Decompression*) em específico, existe uma gama enorme de vários tipos de CODECs que podem ser utilizados para esse fim.

Há inúmeras aplicações que são capazes de reproduzir arquivos nesse formato.

A técnica utilizada para se armazenar os dados nesse formato é chamada de ***Interleaving***. Tal técnica consiste em incorporar dois ou mais dados numa mesma transmissão de informação, conforme Figura 1. Em cada pedaço dessa informação podem ser encontrados dados de vídeo e dados de áudio.

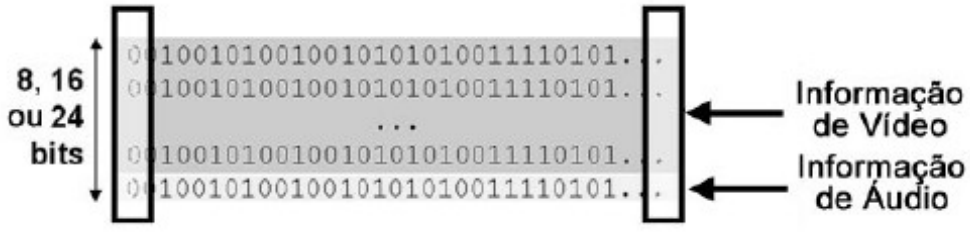


Figura 1: O formato .AVI

CODECS

Vídeos digitais em seu estado bruto requerem muito espaço para serem armazenados. A título de referência, um vídeo em NTSC requer cerca de 27 megabytes a cada segundo para ser armazenado. Devido a esse fato, foram elaborados alguns utilitários para comprimir e descomprimir informações de vídeo e áudio, dessa forma poderia se ter um formato de fácil manipulação, distribuição e armazenamento. É daí que surge a idéia dos ***CODECS***.

CODEC é a abreviação de Compressor/Descompressor e refere-se a qualquer tecnologia implementada em hardware, software ou ainda uma combinação dos dois, utilizada para comprimir e descomprimir dados de vídeo, áudio ou ambos.

Alguns **CODECs** usam uma técnica de compressão sem perda de informação, chamada ***Lossless***. Tal técnica de compressão assegura que toda a informação do vídeo original é preservada após a compressão. Com esse tipo de compressão é possível reconstruir a informação comprimida de modo a restaurar o vídeo em seu formato original. Entretanto, preservar o nível original da qualidade pode limitar o grau que é possível diminuir a taxa de dados e tamanho físico do arquivo do vídeo, dessa forma, o resultado final pode ser muito pior que o desejado para uma reprodução suave do vídeo em alguns computadores.

O formato .TIF

A escolha pelo formato .TIF (*Tagged Image File Format*) se deve, fundamentalmente, ao fato de ser aceita para todas as funções pré-definidas no software MatLab utilizado para teste dos experimentos, além disso outras características desse formato serão mostradas adiante.

O formato .TIF foi desenvolvido em 1986 pela Aldus e pela Microsoft numa tentativa de criar um padrão para imagens geradas por equipamentos digitais. O .TIF é capaz de armazenar imagens *true color* (24 ou 32 bits) e é um formato muito popular para transporte de imagens do desktop para *bureaus*, para saídas de scanners e separação de cores.

O TIF permite que imagens sejam comprimidas usando os métodos

LZW, Packbit, Huffman, CCITT, Group III, Group IV e permite salvar campos informativos (*caption*) dentro do arquivo.

A compressão LZW do formato .TIF, suporta imagens com esquemas de cores não indexadas, como pode ser visto na Figura 3 comparando-a com a Figura 2.

Figura 2: Imagem não compactada – 178KB

Figura 3 Imagem compactada – 157KB (12% de compressão)

O TIF também é muito conhecido e usado para importar/exportar

imagens e fotos entre programas e plataformas (MACS e PCS) diferentes, comprime os arquivos sem perder qualidade da imagem. É muito usado em editoração eletrônica e mídia impressa em geral.

O formato TIF é compatível com a maioria dos sistemas operacionais e é suportado por muitos programas gráficos ou não, contudo não é próprio para uso na web.

Na realidade, os arquivos nos formatos JPEG e GIF possuem tamanhos menores em comparação aos de formatos TIF e BMP. As imagens com formatos TIF, BMP etc., são de alta definição e qualidade, mas têm um tamanho maior. A conversão destes formatos para JPG ou GIF pode ser feita sem dificuldades através de uma grande variedade de programas gratuitos.

O programa

Segue abaixo trechos do código utilizado nos experimentos:

1. **V_c (vetor de cor)** onde a imagem é passada de RGB para HSV, logo após as 3 bandas são separadas, a imagem HSV é quantizada em 5 tons de H, 3 tons de S e 3 tons de V, a nova imagem quantizada é re-montada e finalmente o histograma de cor no espaço HSV com 24 posições é calculado, como a seguir:

```
function FVIm = FVHSV162(Im)

% passa para HSV
Imhsv = rgb2hsv(Im);
// a função rgb2hsv própria do MatLab transforma de RGB para HSV

% separa as três bandas H, S e V
Imh = Imhsv(:,:,1);
Ims = Imhsv(:,:,2);
Imv = Imhsv(:,:,3);

% quantiza a imagem HSV em 18 tons de H, 3 de S e 3 de V
[Imhq,Imsq,Imvq] = QuantizaHSV(Imh,18,Ims,3,Imv,3);

% remonta a imagem quantizada
Imhsv(:,:,1) = Imhq;
Imhsv(:,:,2) = Imsq;
Imhsv(:,:,3) = Imvq;

% calcula o histograma de cor no espaço HSV = 24 posições
FVIm = HistHSV(Imhq,Imsq,Imvq);
```

Para a passagem do espaço RGB para o HSV foi utilizada a função rgb2hsv do *MatLab* e para o cálculo do histograma de cor a função HistHSV, como se segue:

```
function FVIm = HistHSV(Imh,Ims,Imv)

Hh = zeros(1,18);
Hs = zeros(1,3);
Hv = zeros(1,3);

[lin,col] = size(Imh);
for i=1:lin
  for j=1:col

    indh = ceil(Imh(i,j)*18);
    Hh(indh) = Hh(indh) + 1;

    inds = ceil(Ims(i,j)*3);
    Hs(inds) = Hs(inds) + 1;

    indv = ceil(Imv(i,j)*3);
    Hv(indv) = Hv(indv) + 1;

  end
end

FVIm = [Hh Hs Hv];
```

2. **V_e (vetor de forma)** onde é calculado o mapa de bordas e logo após o histograma do mapa de bordas, como a seguir:

```
function FV_EDG_IM = FVEDG10(Im)

% calcula o mapa de bordas
SHAPE = EdgeMapGrad(Im);

% calcula o Histograma do mapa de bordas
```

```
FV_EDG_IM = HistEdge(SHAPE);
```

Para o cálculo do mapa de bordas foi utilizada a função EdgeMapGrad que converte para escala de cinza, depois converte para matriz de índices e finalmente calcula o mapa de bordas, como a seguir:

```
function SHAPE = EdgeMapGrad(Im)

% converte para escala de cinza
if ~isgray(Im)
  Im = rgb2gray(Im);
end

% converte para matriz de índices
Im = double(Im);

% Sobel na Direcao X
Mx = [-1 -2 -1;
      0  0  0;
      1  2  1];

% Sobel na Direcao Y
My = [-1  0  1;
     -2  0  2;
     -1  0  1];

% calcula o ângulo das bordas
[lin,col] = size(Im);
Ime = zeros(lin,col);
for i=1:lin-2
  for j=1:col-2
          Sx = 0;

% calcula o valor do filtro de Sobel horizontal e vertical
Sy = 0;
```

```
      for p=1:3
        for q=1:3
          Sx = Sx + Mx(p,q)*Im(i+p-1,j+q-1);
          Sy = Sy + My(p,q)*Im(i+p-1,j+q-1);
        end
      end

      if Sx == 0
        Ime(i+1,j+1) = 89.999;
      else
       % calcula o ângulo em graus
        Ime(i+1,j+1) = (atan(Sy/Sx))*180/pi;
        if Ime(i+1,j+1) < 0
          Ime(i+1,j+1) = 180 + Ime(i+1,j+1);
        end
      end

    end
  end

  SHAPE = Ime;

  function HShape = HistEdge(SHAPE)

  [lin,col] = size(SHAPE);
  SHAPE = round(SHAPE/19);

  HS = zeros(1,10);
  for i=1:lin
    for j=1:col
      HS(SHAPE(i,j)+1) = HS(SHAPE(i,j)+1) + 1;
    end
  end
  HShape = HS;
```

3. **V_e (vetor de som)** primeiramente o arquivo inteiro de áudio foi convertido para um arquivo binário podendo, dessa forma, passar pela etapa de processamento que consiste em um vetor composto por 7 posições que guardam os valores das características de áudio extraídas, como a seguir:

```
function [Vs] = CaracteristicaSom(s)
Vs=zeros(1,7);
F=Fourier(s);
Ew=AverageEnergy(s);
Ber=BandEnergyRadio(s);
Dsm=DeltaSpectrumMagnitude(s);
Zcr=ZeroCrossingRate2(s);
Rms=RootMeanSquare(s);
Hfvr=HighFeatureValueRatio(s,Zcr);
Lfvr=LowFeatureValueRatio(s,Ew);
Vs=[Ew Ber Dsm Zcr Rms Hfvr Lfvr]
```

```
function Ew=AverageEnergy(s)

[n,m] = size(s);
if n > m  tamanho = n;
else      tamanho = m;
end;

Ew=0;

for i=1:tamanho
   Ew=Ew+s(i)*s(i);
end

Ew;
Ew1 = Ew/tamanho;
```

function Ber=BandEnergyRadio(s,F)

```
resultado1=0;
resultado2=0;

[n,m] = size(s);
if n > m  tamanho = n;
else      tamanho = m;
end;

h = 4; %round(tamanho/4);

for i=1:h
   resultado1=resultado1+F(i);
end
for i=1:8
   resultado2=resultado2+F(i);
end
Ber = resultado1/resultado2;
Ber= sqrt(real(Ber).^2 + imag(Ber).^2);
```

function Dsm=DeltaSpectrumMagnitude(s,F)

```
Dsm=0;
[n,m] = size(s);
if n > m  tamanho = n;
else      tamanho = m;
end;
F= sqrt(real(F).^2 + imag(F).^2);
h=8-1;
for i=1:h
   Dsm=Dsm+abs(abs(F(i))-abs(F(i+1)));
end
```

function Zcr=ZeroCrossingRate2(s)

```
Zcr=0;
[n,m] = size(s);
if n > m  M = n;
else      M = m;
end;
for i=2:M
   if s(i)>=0  arg1=1;
   else        arg1=0; end;

   if s(i-1)>=0  arg2=1;
   else          arg2=0; end;
   Zcr=Zcr+abs(arg1-arg2);
end
Zcr=(Zcr/(M-1));
```

function Rms=RootMeanSquare(s)

```
Rms=0;
M=size(s,1);
for i=1:M
   Rms=Rms + s(i)^2;
end
Rms=sqrt(Rms/M);
```

function Hfvr=HighFeatureValueRatio(s,Zcr)

```
Hfvr=0;
M=size(Zcr,1);
for i=1:M
   Hfvr=Hfvr+abs(sin(Zcr(i)-1.5*mean(Zcr(1:i)))+1);
end
Hfvr=Hfvr/(2*M);
```

function Lfvr=LowFeatureValueRatio(s,Ew)

```
Lfvr=0;
M=size(Ew,1);
for i=1:M
   Lfvr=Lfvr+(sin(0.5*mean(Ew(1:i))-Ew(i))+1);
end
Lfvr=Lfvr/(2*M);
```

function Sim=SimilaridadeSOM(Ve0,Ve1,Vc0,Vc1,Vs0,Vs1)

```
Pc = SimVet (Vc0,Vc1);
Pe = SimVet (Ve0,Ve1);
Ps = SimVet (Vs0,Vs1);
Sim = 1-((1-Pc)*(1-Pe)*(1-Ps));
```

function R = SimVet(p,q)

```
A = sum(p.*q);
B = sqrt(sum(p.*p)) * sqrt(sum(q.*q));
if B ~= 0
   D = A/B;
else
   D = A;
end
R = abs(D);
```

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo