1
Introdução

1.1.
Classificação e Segmentação de Vídeo

O aumento do volume de informações em vídeo armazenadas digitalmente é uma tendência inegável. Cada vez mais aumenta a produção de mídia digital e, adicionando a isso a digitalização de acervos analógicos, cresce muito o tamanho dos bancos de dados multimídia. Assim, técnicas para manipulação de acervos digitais tornam-se extremamente importantes para sua gerência. Essas técnicas, mecanismos de análise automática de vídeo, são focos de muitos estudos recentes, pelo grande ganho de eficiência que representam no trabalho de manipulação e gerência dos vídeos. São elas: segmentação, classificação, indexação e busca de vídeos.

O acesso a vídeos em acervos de larga escala exige técnicas de otimização, que o tornem mais rápido. São muito usados mecanismos de indexação, para que as buscas se tornem mais velozes, facilitando a recuperação dos vídeos. Porém, para que índices sejam criados, é necessário um exame seqüencial do vídeo inteiro, armazenando informações que futuramente possam ser utilizadas como índices para buscas. Assim, a automação desse processo de geração de índices pode representar uma grande economia em força de trabalho para a manutenção de acervos de tamanho considerável.

A classificação e a segmentação são duas técnicas que extraem informações dos vídeos que podem ser utilizadas para a indexação. Na classificação, cada intervalo analisado é classificado pelo tipo de áudio que ocorre. Com isso, é possível identificar tipos de vídeo como um todo, além de trechos de vídeos onde ocorrem mudanças do tipo de áudio predominante. Tanto as informações de classificação de um intervalo quanto as de mudanças do tipo de áudio são muito úteis num processo de indexação, por servirem como uma classificação do vídeo por tipos e indicarem mudanças dessa classificação em determinados momentos de um vídeo – o que muitas vezes indica um ponto que deve ser indexado.

Já na segmentação, o objetivo é buscar regiões semelhantes do vídeo, de modo a dividi-lo em segmentos onde os quadros que estão em cada segmento são relacionados. Podem-se encontrar então desde mudanças de um programa para outro ou mudanças de cena em um programa até mudanças de dinâmica em casos de um clipe musical, representando uma entrada de refrão da música, por exemplo. Assim, pode-se indexar o vídeo por segmentos, tornando muito mais fácil a busca por um trecho específico de um vídeo grande, cuja análise sem o auxílio da automação seria significativamente mais demorada.

Inicialmente, os estudos nessa área focaram-se principalmente nas informações das imagens do vídeo. Porém, cada vez mais se percebe a importância da análise do áudio como um instrumento auxiliar nessa tarefa, aumentando a eficiência e permitindo a descoberta de novas informações. Um exemplo de informação que não pode ser obtida a partir apenas das imagens é a classificação do áudio. Essa informação é muito útil para a análise do tipo de vídeo que está sendo verificado, além de poder ser usada como um mecanismo de auxílio à segmentação. Outros exemplos de informações são: segmentação a partir do áudio, mudança de interlocutor, mudança de volume e detecção de silêncio, todas com utilidade para um processo de análise de vídeos.

1.2.
MPEG-1 Layer 2

1.2.1.
O padrão de áudio MPEG-1 Layer 2

O MPEG é um codificador de áudio que aproveita as limitações do aparelho auditivo humano para remover partes não perceptíveis do áudio e conseguir boas taxas de compressão. Ele é dividido em três camadas, chamadas *layers*: Layer 1, Layer 2 e Layer 3. O Layer 1 é o algoritmo básico, enquanto o Layer 2 e o Layer 3 são extensões dos algoritmos anteriores. Suas capacidades de compressão são maiores, porém isso torna o codificador e o decodificador mais complexos, principalmente no caso do Layer 3.

O áudio MPEG-1 suporta três taxas de amostragem, 32, 44,1 e 48 kHz, e taxas de bits variando desde 32kbps a 448 kbps para o Layer 1, 384 kbps para o Layer 2 e 320 kbps para o Layer 3. É possível a divisão do áudio em um ou dois canais, nos seguintes modos:

- Monofônico: somente um canal de áudio.

- Dual monofônico: dois canais de áudio independentes, como por exemplo o áudio em duas línguas diferentes.

- Stereo: dois canais em estéreo, relacionados, mas com codificação independente.

- Joint-stereo: igual ao modo Stereo; esse modo usa as correlações entre os dois canais para fazer sua codificação, buscando atingir maior qualidade para menores taxas de bits.

Todas as camadas do áudio MPEG-1 utilizam-se de informações do modelo psicoacústico humano para possibilitar altas taxas de compressão. O nosso sistema auditivo tem capacidade limitada, e dependente da freqüência. Isso possibilita fazer a divisão do áudio em sub-bandas, em vez de se representar cada freqüência individualmente. Sabe-se também que o domínio de freqüências que um ouvido humano consegue perceber está entre 20 Hz e 20 kHz, e a faixa de maior sensibilidade está entre 2 e 4 kHz. O domínio de volumes escutado, do silêncio ao mais alto som, é outra informação utilizada. Ele é de aproximadamente 96 dB.

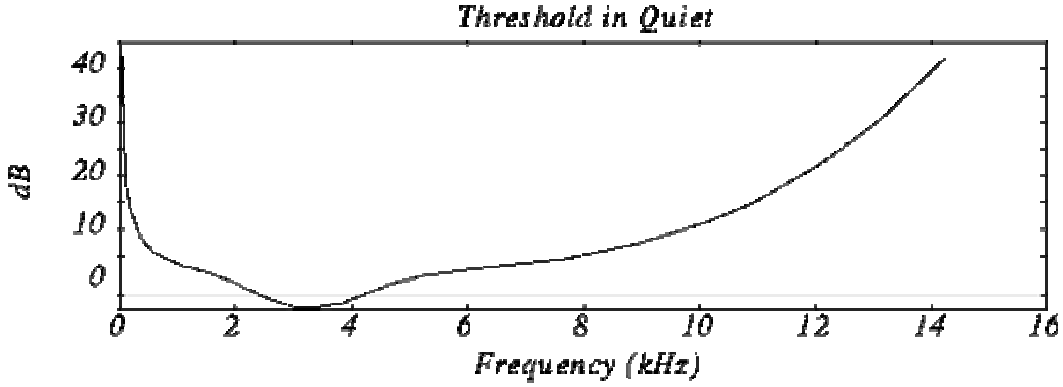

Figura 1 – Sensibilidade do ouvido humano

Outra característica explorada pelo MPEG-1 é o efeito de mascaramento. Ele se divide entre mascaramento freqüêncial e temporal. O mascaramento freqüêncial é empregado em todos os *layers* do MPEG-1. Quando dois sinais de freqüências muito próximas ocorrem, se houver uma diferença considerável de volume o sinal de maior volume mascara o outro sinal; isto é, o ouvido humano não tem capacidade de discerni-lo. Esse fato é aproveitado pelo MPEG-1 para diminuir a quantidade de informação registrada e possibilitar o uso de sub-bandas. Já o mascaramento temporal, presente apenas no Layer 3, se dá quando um sinal de alto volume é seguido por um de volume mais baixo em uma freqüência próxima. Durante um período de tempo, o segundo sinal não pode ser detectado pelo ouvido humano.

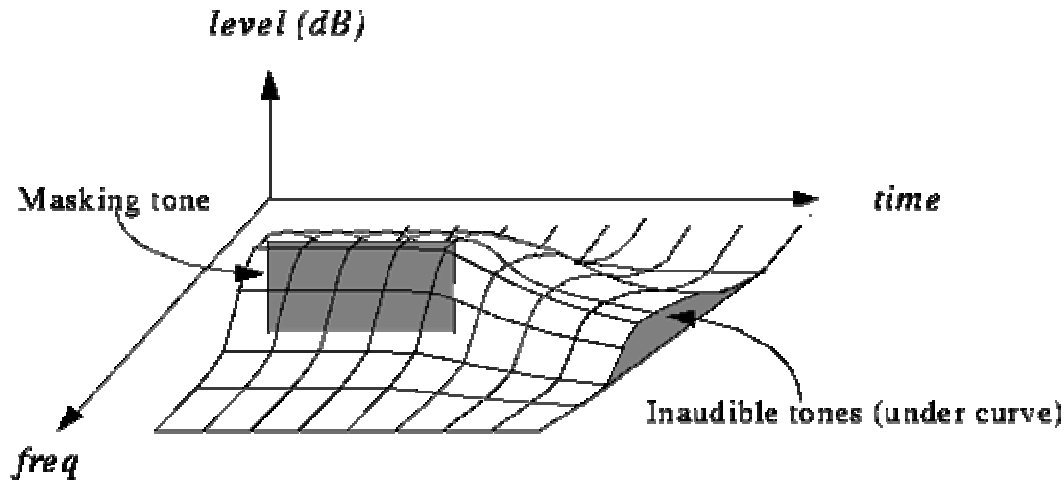

Figura 2 – Mascaramento de freqüências e temporal

Nos *layers* 1 e 2, o áudio é dividido em 32 sub-bandas de freqüências com tamanhos iguais, sendo armazenado um valor de amostra para cada sub-banda.

Cada sub-banda cobre uma faixa de aproximadamente 650 Hz do espectro, possibilitando com que todas as freqüências que o ouvido humano tem capacidade de escutar sejam armazenadas. As sub-bandas mais baixas se localizam nas freqüências menores, de forma que quanto maior a sub-banda, maior a freqüência do áudio que ela representa. A codificação por sub-bandas se baseia na idéia de mascaramento de freqüências, de forma que a diferença no sinal por causa das informações descartada não seja notada. No Layer 1 temos blocos de 12 amostras para cada sub-banda, totalizando 384 amostras por *frame*. Já no Layer 2 temos 3 grupos de blocos de 12 amostras, totalizando 1152 amostras por *frame*.

Um quadro de áudio possui, além da informação do áudio, também um cabeçalho (*Header*), um campo opcional de verificação de erros por redundância cíclica (*CRC*) e dados auxiliares (*Ancillary Data*). A Figura 3 ilustra o formato dos quadros de áudio MPEG:

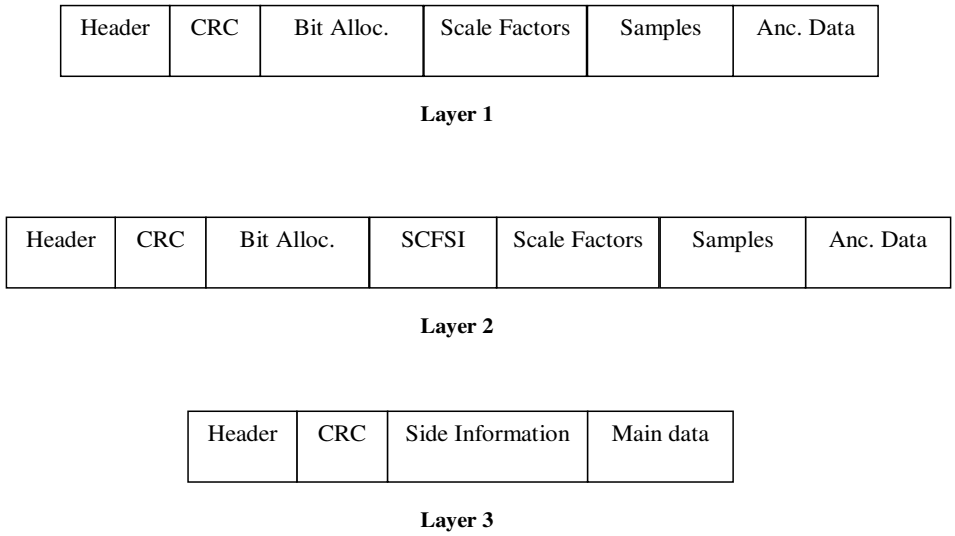

Figura 3 – Formato dos quadros de áudio MPEG

A informação do áudio em uma sub-banda do Layer 1 é composta de três informações:

- *Bit allocation*: indica o número de bits utilizado na codificação de cada amostra.

- *Scale Factor*: é um multiplicador que redimensiona as amostras, permitindo o uso do alcance completo do decodificador. Indica o volume máximo de um grupo de amostras da sub-banda.

- *Sub-band samples*: são as amostras em si, que serão recodificadas a partir dos *scale factors* para reconstruir o valor original e posteriormente a onda sonora.

No Layer 2 há um campo adicional, contendo as informações de seleção dos *scale factors* (*SCFSI*). Esse campo indica como os *scale factors* são compartilhados, permitindo que sejam transmitidos em maior quantidade caso mudanças significativas ocorram. Nesse caso, é necessário um maior detalhamento das informações do áudio, e a presença de *scale factors* diferentes onde só haveria um no Layer 1 permite que isso ocorra.

O Layer 3 combina algumas características dos *layers* 1 e 2 com uma codificação adicional. É aplicada uma transformada de cossenos (*MDCT*), além da codificação de Huffman. O uso da transformada de cossenos permite a divisão de cada sub-banda em 18 sub-bandas menores. Isso possibilita uma maior precisão para taxas de bit menores, porém os processos de codificação e decodificação se tornam mais complexos e trabalhosos. O Layer 3 foge do escopo desta dissertação, todavia mais detalhes sobre esse padrão podem ser encontrados em ISO (1993b).

1.2.2. Motivos da escolha do padrão

Acervos digitais normalmente possuem dois tipos de cópias do mesmo vídeo: uma cópia de preservação e uma cópia de gerência. A cópia de preservação é armazenada de forma a não perder informações, preservando o vídeo original. Já a cópia de gerência é normalmente armazenada comprimida, pois a qualidade do vídeo não precisa ser preservada. Então é vantajoso nesse domínio trabalhar com vídeos de menor tamanho possível, que se adequem ao sistema de gerenciamento utilizado. Isso diminui a capacidade de armazenamento exigida dos bancos de dados, e facilita a manipulação dessas cópias.

Nesse contexto, o MPEG-1 é um excelente padrão. Ele possui uma boa taxa de compressão, sem incluir complexidade demais no codificador e no decodificador, o que o torna bastante eficiente. Os sistemas de gerência trabalham com um padrão específico para o qual são desenvolvidos. Sendo assim, as possibilidades apresentadas ao se trabalhar com esse padrão em vez de procurar o

desenvolvimento de um algoritmo genérico, que certamente seria mais lento e ineficiente, tornam essa escolha compensadora.

Dentre as camadas de áudio do MPEG-1, o áudio Layer 2 é o que melhor se encaixa para este trabalho. Seu processo de decodificação, menos complexo que o do Layer 3, sua maior resistência a erros e seu desempenho, que em altas taxas de bits é melhor do que o do Layer 3, tornaram o Layer 2 o padrão para aplicações *broadcast*. Além disso, ele se torna extremamente vantajoso ao se trabalhar no domínio comprimido.

A análise dos vídeos nesse domínio oferece um grande ganho de eficiência. O fato de não ser necessário descomprimir o vídeo para que sejam aplicadas as técnicas de análise faz com que a rapidez desse processo seja muito maior. Como tratamos de acervos em larga escala, esse ganho de eficiência se torna um importante diferencial.

O Layer 2 permite que a análise do vídeo seja feita com a leitura do mínimo de informações. Isso porque os *scale factors* armazenados representam uma indicação do volume máximo das suas sub-bandas. Assim, é possível extrair informações a partir apenas dos *scale factors*, sem ser necessária a leitura dos valores dos *samples*. Isso representa um ganho de desempenho significativo, pois apesar de se trabalhar com informações menos detalhadas, é necessário ler apenas uma fração das informações armazenadas. Além disso, combinando seus resultados com a análise das imagens, é possível atingir um alto grau de precisão.

Um último fator a ser considerado é o pequeno número de resultados apresentados em pesquisas realizadas a partir apenas dos *scale factors*. Esse assunto será detalhado no capítulo 2, porém poucos resultados significativos foram encontrados, o que torna a técnica apresentada nesta dissertação ainda mais significativa.

1.3. Escopo da dissertação

Esta dissertação se foca nas tarefas de classificação e segmentação de áudio no domínio comprimido, utilizando o padrão MPEG-1 Layer 2. A análise é realizada a partir apenas dos *scale factors* dos *frames* de áudio, assim é sacrificado o nível de detalhes das informações obtidas para a análise em troca de um grande ganho de desempenho.

O algoritmo proposto faz a classificação do áudio em quatro tipos (silêncio, fala, música e aplauso) e usa essa informação também para auxiliar a tarefa de segmentação. O nível de precisão atingido pelo algoritmo de classificação proposto é muito alto, e ele apresenta a detecção de aplausos como uma novidade em relação à literatura. A segmentação a partir apenas dos *scale factors* é outra inovação, pois não foi encontrado trabalho semelhante nas pesquisas realizadas.

1.4. Organização da Dissertação

Este capítulo apresentou uma introdução ao tema e ao trabalho, além de um resumo breve do padrão MPEG-1 para áudio. O capítulo 2 apresenta um levantamento dos trabalhos relacionados encontrados na literatura. Os capítulos 3 e 4 apresentam o algoritmo proposto para a classificação e a segmentação, respectivamente. Os resultados obtidos por esses algoritmos são mostrados no capítulo 5. O capítulo 6 apresenta as conclusões e sugestões de trabalhos futuros.