1
Introdução

1.1.
Indexação, Recuperação e Segmentação de Vídeo

Os acervos volumosos de vídeo estão se popularizando motivados pelo avanço das tecnologias de captura, armazenamento e compressão de vídeo digital, pela significativa melhoria da capacidade de transmissão e recepção de redes com acesso doméstico e pelo aumento da capacidade de processamento e armazenamento dos computadores. Esses fatores motivam o desenvolvimento e a disponibilização de novos serviços e produtos para manipulação e gerenciamento de acervos de vídeo digital.

Entre os produtos diretamente relacionados a acervos de vídeo destacam-se a televisão interativa, sistemas multimídia, bibliotecas digitais de vídeo, serviços de telemedicina, ambientes virtuais de aprendizado, televisão digital e serviços de transmissão de vídeo através de redes com acesso doméstico.

Um novo serviço sendo amplamente pesquisado é o fornecimento de vídeo por demanda (Tobin, 1999). Esse serviço permite que acervos de arquivos de vídeo digital, com conteúdos variados, como filmes ou programas de TV aberta, possam ser remotamente acessados e armazenados em servidores com alta capacidade de armazenamento e transmissão.

A obtenção de um vídeo em um acervo de massa, entretanto, requer mecanismos de indexação que permitam o acesso rápido ao seu conteúdo. A geração manual de índices está vinculada ao exame seqüencial de todo o vídeo. Esse processo é extremamente oneroso pelo volume de dados manipulados, inviabilizando sua utilização em grandes acervos. A automação desse processo de indexação parte da geração de granularidades mais finas de acesso ao vídeo, permitindo visualizar seu conteúdo sem que seja necessário exibir todo o vídeo. Este processo pode ser feito com o auxílio da segmentação automática (ou semi-automática) de vídeo.

De um ponto de vista bastante geral, segmentação de vídeo se refere à identificação de regiões em um quadro de vídeo que são homogêneos em algum sentido (Tekalp, 2000). O principal propósito da segmentação de vídeo é possibilitar uma representação baseada em conteúdo através da extração de objetos de interesse a partir de uma série de quadros consecutivos de vídeo – um tópico essencial em visão computacional. Há vários tipos de segmentação, tais como as segmentações de cor, textura e movimento. Neste trabalho, a segmentação de vídeo tem a conotação específica de se referir à detecção de transição entre tomadas de câmera, como um corte seco ou uma transição gradual de uma cena para outra. Do ponto de vista teórico, esse processo de segmentação baseia-se em medidas de diferença entre quadros como uma forma de definir similaridade. A segmentação de tomadas de câmera é essencial para indexação, recuperação e navegação (*browsing*) em acervos de vídeo.

O conteúdo de um vídeo possui informações muito complexas para serem tratadas pelas funções usuais de consulta. Existem métodos de busca denominados consulta via objeto de conteúdo (*query-by-content-object)*, nos quais são especificadas relações temporais e espaciais entre os objetos procurados no vídeo (Chen et al., 1999). Entretanto, os métodos conhecidos para reconhecimento de objetos no interior de um vídeo são computacionalmente custosos e os índices para os objetos reconhecidos não são construídos automaticamente.

Os métodos de recuperação de um vídeo baseados em características das seqüências de imagens do vídeo são mais eficientes e práticos de serem computados. Nesse caso, a organização do acervo é feita através de mecanismos para a medição da similaridade entre os quadros. A partir da análise de funções de semelhança, são construídas estruturas cinemáticas hierárquicas, ou seja, estruturas hierárquicas baseadas na mecânica do movimento através do tempo (Michaelis, 1999). Essa construção é feita como passo inicial do tratamento semântico do vídeo, permitindo que os níveis da hierarquia sejam utilizados como unidades de manipulação e recuperação do vídeo. Quadros-chave representativos dessas unidades podem ser extraídos como índices de busca (Chen et al., 1999; Ngo et al., 1998).

Os níveis da hierarquia apresentam graus diferentes de abstração semântica. Um exemplo de estrutura cinemática hierárquica é a divisão nos seguintes níveis (Lew et al., 2000): tomada de câmera, cena e episódio (ver Figura 1). O nível mais

baixo dessa hierarquia é o mais próximo das características físicas do vídeo, sendo o primeiro nível a ser identificado pelos processos de segmentação. Nesse nível a unidade utilizada é composta por uma seqüência de quadros no intervalo formado por uma tomada de câmera.

Uma tomada de câmera é composta pela seqüência de quadros contínuos temporal e espacialmente, ou seja, imagens formadas por uma gravação ininterrupta de câmera (Tobin, 1999). Na literatura sobre o assunto é comum a denominação imprópria de mudança de cena para classificar a fronteira entre duas tomadas de câmera. Entretanto, seguindo a nomenclatura da indústria cinematográfica, uma cena é uma coleção de tomadas de câmera agrupadas por pertencerem a um ato contínuo de apresentação em um mesmo ambiente e, portanto, forma o nível seguinte da estrutura hierárquica.

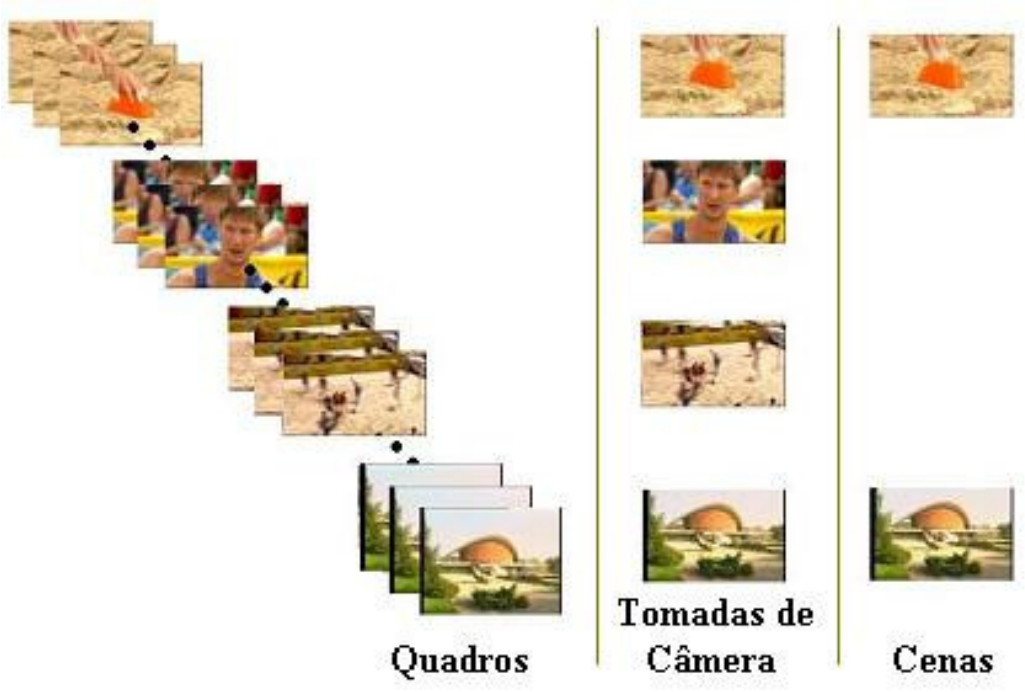

Figura 1 – Estrutura Cinemática Hierárquica

1.2. Os Padrões MPEG-1 e MPEG-2

Para o desenvolvimento desta dissertação sobre segmentação de vídeo em tomadas de câmera, faz-se necessária a adoção de um padrão de vídeo digital. Em médio prazo, o padrão MPEG-2 é apontado como sendo o padrão de vídeo digital com maior abrangência entre fabricantes de hardware e desenvolvedores de

aplicações para vídeo digital (Strachan, 1996). Conseqüentemente, optar pela compatibilidade com os padrões MPEG-1 e MPEG-2 como requisito inicial, permite abranger um grupo significativo de consumidores. Em especial destaque entre as aplicações compatíveis com o padrão MPEG-2, encontram-se os sistemas de transmissão terrestre de televisão digital europeu, Digital Video Broadcasting (DVB) e o padrão norte-americano, US Grand Alliance HDTV (High Definition Television). Por estas razões, o presente trabalho adota os padrões MPEG-1 e MPEG-2. As referências aos padrões MPEG-1 e MPEG-2 no desenvolvimento deste texto serão mencionados, apenas, como padrão MPEG.

1.3. Escopo da Dissertação

Esta dissertação propõe um algoritmo para a segmentação de vídeos MPEG-1 e MPEG-2 em unidades de tomada de câmera. O problema da detecção das tomadas de câmera de um vídeo é equivalente ao problema identificação dos quadros posicionados na fronteira entre duas tomadas de câmera consecutivas. Em especial, denomina-se o primeiro quadro de uma tomada de câmera como quadro-corte. Sendo assim, o intervalo entre dois quadros-corte estabelece as margens e a duração de uma tomada de câmera. As operações de edição e composição das tomadas de câmera para a montagem dos vídeos fornecem diversas formas de transição entre as tomadas de câmera. As formas mais utilizadas para composição da transição entre duas tomadas de câmera são o corte seco, o *dissolve*, o *fade-in,* o *fade-out* e o wipe. Esta dissertação está focalizada no desenvolvimento de um algoritmo para detecção de cortes secos. Os cortes graduais são detectados, mas não são classificados. Para sua classificação, faz-se necessária a ampliação do algoritmo proposto para também incorporar heurísticas de classificação dos tipos de transição.

Os diversos tipos de algoritmo para detecção das fronteiras entre duas tomadas de câmera se diferenciam principalmente pela métrica usada para identificação dos quadros-corte e pela natureza dos dados analisados por essas métricas.

A comparação entre as diversas abordagens ressalta os critérios de eficiência em relação à velocidade de processamento, distinguindo métricas que manipulam

os dados do fluxo de dados (*stream*) codificados de métricas aplicadas sobre as características espaciais das imagens. Esse segundo grupo apresenta um pior desempenho, relacionado aos fatores diferenciais do alto custo de descompressão dos dados do vídeo e do grande volume de dados operados no domínio descomprimido.

Pelos motivos apresentados acima, o algoritmo proposto nesta dissertação possui o pré-requisito de analisar apenas informações obtidas diretamente dos dados do vídeo codificado, eliminando as etapas de processamento da descompressão, garantindo a redução do volume de dados processado. Também por questões de desempenho, o algoritmo foi modelado para fornecer análises hierárquicas e assim permitir o descarte simultâneo de grupos de quadros.

A etapa de maior complexidade, entre as diversas abordagens para identificação dos cortes, é a etapa de seleção de um limiar para classificação dos quadros dos vídeos pela métrica adotada (Ceccarelli et al., 1997). Essa complexidade justifica-se pelo fato de não ser conhecida uma função de distribuição ideal para o problema da detecção de tomadas de câmera. Não é conhecido um critério de análise para o qual exista pelo menos um limiar separando perfeitamente os quadros-corte dos quadros não-corte.

As limitações do uso de uma única métrica para a classificação de todo o vídeo, relacionadas à possível inexistência de um limiar ideal, motivam a nova abordagem proposta no presente trabalho. Usando variadas análises para caracterização dos quadros, as funções características são aplicadas com o intuito de eliminar a possibilidade de corte em quadros que apresentem semelhança a quadros anteriores na seqüência de vídeo. O relaxamento do limiar é compensado pela aplicação de etapas sucessivas de refinamento por critérios variados.

Também como contribuição inédita ao problema da detecção das tomadas de câmera, a abordagem proposta nesta dissertação identifica e caracteriza padrões falsos de cortes nas distribuições das métricas de análise dos dados comprimidos. A existência destes falsos padrões está associada às escolhas e aos parâmetros do processo de compressão. Os padrões MPEG-1 e MPEG-2 não estabelecem a implementação fechada de algoritmos para atender aos seus requisitos de compressão, de forma a existir uma liberdade na definição dos procedimentos e, conseqüentemente, permitindo que diferentes implementações de compactadores gerem fluxos de dados (*streams*) distintos para a codificação de um mesmo vídeo.

Nesta dissertação, as escolhas livres da aplicação de compressão são denominadas "inteligência do codificador".

A identificação desses falsos padrões é feita pelo levantamento da "história" de codificação através de uma análise comparativa do comportamento dos quadros de todo o vídeo. Os dados para esse levantamento são extraídos das características dos vídeos codificados. Os padrões falsos de corte detectados são chamados de "assinatura" do codificador.

Para distinguir os quadros com reais características de corte dos quadros com características influenciadas pelo codificador (ou seja, evitar erros de detecção associados aos padrões intrínsecos às escolhas de codificação), são propostas filtragens nos dados analisados, de forma a suavizar a influência das assinaturas de codificação nos valores obtidos pelas métricas de caracterização de similaridade.

Nenhuma característica extra aos padrões de codificação MPEG-1 e MPEG-2 é assumida pelo algoritmo proposto, assegurando sua generalização para vídeos comprimidos em qualquer codificador aprovado pelas normas desses padrões. Para garantir a flexibilidade do algoritmo, são fornecidos dois parâmetros de adaptação: parâmetro de sensibilidade e parâmetro de vizinhança mínima.

Inserido no contexto atual de pesquisas em gerenciamento de acervos de vídeo em busca por soluções para disponibilizar seu conteúdo de maneira eficiente, este estudo trata o seguinte:

- o problema da segmentação de vídeo em unidades de tomada de câmera em domínio compactado, mais especificamente, vídeo digital padrão MPEG-1 e MPEG-2;
- as abordagens encontradas na literatura; e
- o algoritmo desenvolvido para o tratamento desse problema.

1.4. A Organização da Dissertação

O levantamento geral dos trabalhos relacionados ao problema da segmentação de vídeo digital encontrados na literatura é exposto no Capítulo 2 desta dissertação. O Capítulo 3 apresenta o algoritmo desenvolvido. A análise dos resultados obtidos pela aplicação do algoritmo proposto ao acervo de vídeo criado

para teste é apresentada no Capítulo 4. O Capítulo 5 apresenta as conclusões finais e sugestões de trabalhos futuros.

Para uma melhor compreensão do trabalho por leitores leigos em vídeo MPEG-1 e MPEG-2, são inseridos três anexos, que devem ser consultados por esses leitores antes da leitura dos demais capítulos. Os apêndices estão organizados da seguinte forma: o Apêndice I apresenta os padrões MPEG-1 e MPEG-2; o Apêndice II apresenta a estrutura hierárquica de codificação do vídeo utilizada por esses padrões; e o Apêndice III apresenta a compressão MPEG.