Recuperação de Vídeos Comprimidos por Conteúdo*

Jurandy Almeida, Neucimar J. Leite, Ricardo da S. Torres

Instituto de Computação, Universidade Estadual de Campinas – UNICAMP
13083-852, Campinas, SP – Brasil
{jurandy.almeida,neucimar,rtorres}@ic.unicamp.br

Abstract. *Most of existing systems for video retrieval rely on algorithms and methods which are computationally expensive, limiting their scope of application. Contrary to this trend, the market has shown a growing demand for mobile and embedded devices. In this context, this thesis introduces five novel approaches for the analysis, indexing, and retrieval of digital videos in limited capacity devices. All these contributions are combined to create a computationally fast system, which is able to achieve a quality level superior to current solutions.*

Resumo. *A maioria dos sistemas existentes para a recuperação de vídeos envolve algoritmos e métodos computacionalmente custosos, limitando o seu escopo de aplicação. Contrário a essa tendência, o mercado tem mostrado uma crescente demanda por dispositivos móveis e embutidos. Nesse contexto, esta tese introduz cinco abordagens originais voltadas a análise, indexação e recuperação de vídeos digitais em dispositivos de capacidade limitada. Todas essas contribuições são somadas na criação de um sistema computacionalmente rápido, capaz de atingir a um padrão de qualidade superior a soluções atuais.*

1. Introdução

Avanços recentes na tecnologia têm permitido o aumento da disponibilidade de dados de vídeo, criando grandes coleções de vídeo digital. Isso tem despertado grande interesse em sistemas capazes de gerenciar esses dados de forma eficiente. Fazer uso eficiente de informações de vídeo requer o desenvolvimento de sistemas capazes de abstrair representações semânticas de alto nível a partir das informações de baixo nível dos vídeos, conhecidos por sistemas de gestão de vídeos por conteúdo [Almeida 2011].

A Figura 1 mostra a arquitetura básica desses sistemas. Devido à complexidade do material de vídeo, existem cinco desafios principais na construção de tais sistemas [Almeida 2011]: (1) dividir o fluxo de vídeo em trechos manuseáveis de acordo com a sua estrutura de organização, (2) implementar algoritmos para codificar as propriedades de baixo nível de um trecho de vídeo em vetores de características, (3) desenvolver medidas de similaridade para comparar esses trechos a partir de seus vetores, (4) responder rapidamente a consultas por similaridade sobre uma enorme quantidade de sequências de vídeo e (5) apresentar os resultados de forma amigável ao usuário.

Inúmeras técnicas têm sido propostas para atender a tais requisitos. A maioria dos trabalhos existentes envolve algoritmos e métodos computacionalmente custosos, em termos tanto de tempo quanto de espaço, limitando a sua aplicação apenas ao ambiente acadêmico e/ou a grandes empresas. Contrário a essa tendência, o mercado tem mostrado uma crescente demanda por dispositivos móveis e embutidos. Nesse cenário, é imperativo

*Os autores agradecem o apoio financeiro das agências FAPESP, CNPq e CAPES.

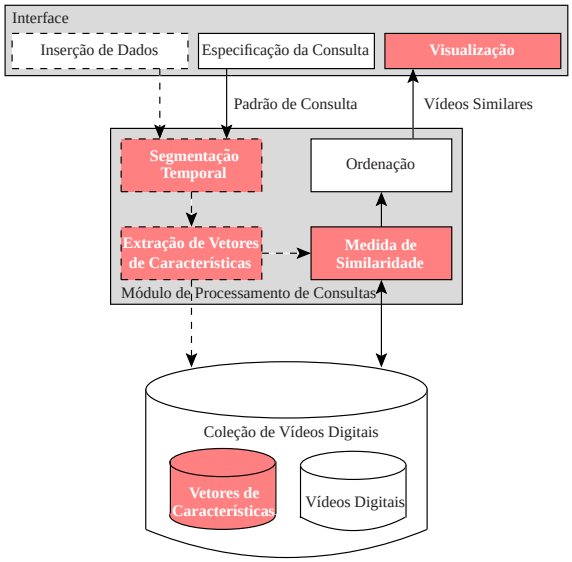

Figura 1. Arquitetura típica de um sistema de gestão de vídeos por conteúdo.

o desenvolvimento de técnicas tanto eficazes quanto eficientes a fim de permitir que um público maior tenha acesso a tecnologias vigentes.

O trabalho desenvolvido nesta tese contribuiu para resolver esse problema. Nesse contexto, o objetivo da tese foi oferecer soluções a vários problemas em aberto na literatura de processamento de vídeos, voltados à análise, indexação e recuperação de vídeos digitais; e contribuir com a superação dos desafios de pesquisa envolvidos na especificação e implementação de sistemas de gestão de vídeos por conteúdo que possam ser aplicados a dispositivos com baixo poder computacional e/ou em ambientes que necessitem de uma resposta rápida, mantendo um padrão de qualidade comparável ao estado da arte.

A principal contribuição da tese foi a introdução de cinco abordagens originais, uma para cada módulo da arquitetura básica de um sistema de gestão de vídeos por conteúdo, os quais estão destacados na Figura 1. Todas essas técnicas foram projetadas para serem, ao mesmo tempo, eficientes e eficazes, a fim de torná-las escaláveis e, portanto, adequadas a grandes coleções de vídeos, podendo ser aplicadas com sucesso em dispositivos com capacidade limitada. Elas são [Almeida 2011]:

1. Uma nova abordagem para a segmentação temporal de sequências de vídeo (Segmentação Temporal). A eficiência computacional dessa técnica a torna adequada para tarefas *online* [Almeida et al. 2011b].
2. Uma nova abordagem para representar o movimento da câmera em sequências de vídeo (Extração de Vetores de Características). Essa técnica relaciona os parâmetros de movimento diretamente com as operações físicas da câmera [Almeida et al. 2009].
3. Uma nova abordagem para a comparação de sequências de vídeo (Medida de Similaridade). A eficiência computacional dessa técnica a torna adequada a grandes coleções de vídeo [Almeida et al. 2011a].
4. Uma nova estrutura de indexação para realizar de buscas por similaridade em espaços métricos (Vetores de Características). Essa técnica é escalável, o que a torna adequada a grandes volumes de dados [Almeida et al. 2010b].
5. Uma nova abordagem para a sumarização de sequências de vídeo, que permite a customização do usuário (Visualização). Essa técnica foi projetada para a

produção de resumos de vídeo tanto estáticos quanto dinâmicos em tarefas *online* [Almeida et al. 2010c, Almeida et al. 2012b, Almeida et al. 2012a].

Uma outra contribuição foi a avaliação empírica dessas técnicas contra métodos clássicos em análise, indexação e recuperação de vídeos digitais. Para isso, diversos experimentos foram conduzidos em várias coleções de vídeo, usando protocolos experimentais imparciais, controlados e reproduzíveis. Todos esses experimentos foram cuidadosamente projetados para assegurar significância estatística, permitindo discernir diferenças genuínas em desempenho de diferenças casuais.

Por fim, todas essas contribuições foram somadas na construção de um sistema de gestão de vídeos por conteúdo computacionalmente rápido, capaz de atingir a um padrão de qualidade superior a soluções atuais. Além das contribuições mencionadas, muitas outras derivadas desta tese foram publicadas em [Almeida et al. 2010a, Almeida et al. 2010d, Pinto-Cáceres et al. 2011, Kozievitch et al. 2011b, Li et al. 2011, Kozievitch et al. 2011a]. As seções a seguir detalham cada uma das contribuições supracitadas, provenientes da pesquisa desenvolvida nesta tese.

2. Segmentação Temporal

O primeiro passo para gerenciar a informação de vídeo é estruturá-la em um conjunto de unidades compreensíveis e gerenciáveis, de forma que o seu conteúdo seja consistente em termos de operações da câmera e eventos visuais. Isso tem sido o objetivo de uma área de pesquisa bastante conhecida, denominada segmentação de vídeo.

Diferentes técnicas têm sido propostas na literatura para lidar com a segmentação temporal de sequências de vídeo. Muitos desses trabalhos de pesquisa estão centrados no domínio descomprimido. Embora os métodos existentes proporcionem uma qualidade elevada, a decodificação e a análise de uma sequência de vídeo são duas tarefas extremamente lentas e requerem uma enorme quantidade de espaço.

A contribuição publicada em [Almeida et al. 2011b] introduz uma nova abordagem para a segmentação temporal de sequências de vídeo, que opera diretamente no domínio comprimido. Ela baseia-se na exploração de características visuais extraídas do fluxo de vídeo e em um algoritmo simples e rápido para detectar mudanças temporais. A melhoria da eficiência computacional torna essa técnica adequada para tarefas *online*.

O algoritmo proposto foi avaliado em um conjunto de testes com diferentes gêneros de vídeo e comparado com abordagens populares na literatura de segmentação temporal. Resultados de uma avaliação experimental sobre vários tipos de transições entre tomadas mostram que esse método é, ao mesmo tempo, bastante preciso e rápido.

3. Representação do Conteúdo Visual

Fazer uso eficiente da informação de vídeo requer que os dados sejam armazenados de maneira compacta. Para isso, um vídeo deve ser associado a características apropriadas, a fim de permitir qualquer recuperação futura. Uma característica importante é a mudança de intensidade temporal entre quadros sucessivos, a qual é geralmente atribuída ao movimento causado por objetos ou introduzido por operações de câmera.

Inúmeros algoritmos têm sido propostos na literatura para estimar o movimento da câmera a partir das sequências de vídeo. Essas soluções consistem tipicamente de uma abordagem de duas etapas: primeiro identificando o movimento e em seguida associando-o a um modelo paramétrico. A forma paramétrica mais popular utilizada é o modelo afim, o qual não está diretamente relacionado às operações físicas da câmera.

A contribuição publicada em [Almeida et al. 2009] introduz uma nova abordagem para estimar o movimento da câmera em sequências de vídeo, com base em um modelo de operações de câmera. Esse método gera o modelo de câmera usando combinações lineares de protótipos de fluxo ótico produzidos por cada operação de câmera.

Para a avaliação dessa técnica, foi utilizado um conjunto de teste sintético e sequências de vídeo reais, incluindo todas as operações básicas de câmera e muitas de suas combinações possíveis. Além disso, foram realizados vários experimentos para mostrar que essa técnica é mais eficaz do que as abordagens baseadas no modelo afim.

4. Medida de Similaridade

Após obter representações compactas, é preciso ainda definir uma medida de similaridade para comparar vídeos a partir de suas assinaturas. Existem duas questões importantes nessa tarefa: robustez e discriminância. Robustez é a quantidade de inconsistência dos dados tolerada pelo sistema antes da ocorrência de um falso positivo. Discriminância é a capacidade do sistema de rejeitar dados irrelevantes e reduzir os falsos positivos.

Diferentes técnicas têm sido propostas na literatura para resolver o problema da comparação entre sequências de vídeo. Embora os métodos existentes proporcionem uma qualidade elevada em termos de robustez e discriminância, a principal desvantagem deles é que a assinatura gerada é proibitiva em termos de espaço de armazenamento, e sua comparação usando uma medida de similaridade baseada em uma abordagem quadro a quadro é impraticável em bases de dados muito grandes.

A contribuição publicada em [Almeida et al. 2011a] introduz uma nova abordagem para a comparação de sequências de vídeo, que atua diretamente no domínio comprimido. Ela baseia-se no reconhecimento de padrões de movimento extraídos do fluxo de vídeo, que são acumulados para formar um histograma normalizado. Essa abordagem computacionalmente simples é robusta a várias distorções e transformações. A melhora da eficiência computacional torna essa técnica adequada a enormes coleções de vídeo.

O algoritmo proposto foi avaliado em cerca de 11.500 vídeos (400 horas) de um conjunto de dados do TRECVID 2010 (IACC.1) e comparado com abordagens recentes na literatura de detecção de similaridade entre vídeos. Resultados de uma avaliação experimental sobre vários tipos de transformações de vídeo mostram que esse método apresenta alta eficácia e velocidade computacional na identificação de vídeos similares.

5. Indexação de Dados

Quando um usuário especifica um padrão de consulta ao sistema, a sua assinatura é extraída e a medida de similaridade é aplicada para identificar todas as assinaturas similares. Para garantir uma resposta rápida, é necessário o desenvolvimento de algoritmos que acelerem esse processo. Elaboradas estruturas de indexação têm sido propostas a fim de organizar as assinaturas extraídas e facilitar a busca por vídeos semelhantes.

A maioria dos índices existentes é construída a partir do particionamento de um conjunto de objetos usando somente a informação de distância entre eles. A fim de manter o balanceamento da estrutura, tal conjunto é fragmentado em partes de mesmo tamanho, ignorando o agrupamento inerente de seus objetos. Em geral, essas técnicas podem ser divididas em duas categorias diferentes. Uma classe de métodos produz partições, portanto, essas partes são disjuntas, o que pode separar objetos próximos, afetando gravemente a eficiência de busca. Outra classe de métodos produz agrupamentos, portanto, essas partes podem se sobrepor, o que pode degradar consideravelmente o tempo das consultas.

A contribuição publicada em [Almeida et al. 2010b] avalia o desempenho de uma nova estrutura de indexação, chamada de BP-tree (do inglês, *Ball-and-Plane tree*), a qual é construída dividindo-se um conjunto de objetos em grupos compactos. Ela combina vantagens de ambos os paradigmas disjuntos (partições) e não disjuntos (agrupamentos) a fim de obter uma estrutura de aglomerados densos e pouco sobrepostos, melhorando a eficiência no processamento de consultas por similaridade.

Essas propriedades da BP-tree são apoiadas por uma extensa avaliação experimental realizada sobre várias bases de dados. Os resultados demonstram que essa abordagem supera as soluções tradicionais. Além disso, ela é escalável, exibindo um comportamento sublinear em relação ao número de objetos, o que a torna adequada a grandes coleções.

6. Visualização dos Resultados

No final, os usuários são apresentados com uma lista de vídeos similares a um dado padrão de consulta especificado. É inviável esperar que um usuário assista a todo o conteúdo desses vídeos ou uma boa parte dele, a fim de descobrir do que eles realmente se tratam. Portanto, é importante fornecer aos usuários uma representação concisa do vídeo para dar uma ideia do seu conteúdo, sem ter que vê-lo completamente, de modo que um usuário possa decidir se quer ou não assistir ao vídeo todo.

Diferentes técnicas têm sido propostas na literatura para resolver o problema de resumir sequências de vídeo. Muitos desses trabalhos de pesquisa estão centrados no domínio descomprimido. Devido ao longo tempo gasto para decodificar e analisar uma sequência de vídeo, os resumos são muitas vezes produzidos completamente *offline*, armazenados e entregues a um usuário quando solicitado. O inconveniente dessa abordagem é a completa falta de personalização pelo usuário.

A contribuição publicada em [Almeida et al. 2010c, Almeida et al. 2012b, Almeida et al. 2012a] introduz VISON[1] (do inglês, *VIdeo Summarization for ONline applications*), uma nova abordagem para resumir sequências de vídeo, que atua diretamente no domínio comprimido. Ela baseia-se na exploração de características visuais extraídas do fluxo de vídeo e em um algoritmo simples e rápido para resumir o seu conteúdo. A melhoria da eficiência computacional torna essa técnica adequada para tarefas *online*. Tal método foi projetado para oferecer customização: usuários podem controlar a qualidade dos resumos e também especificar o tempo que estão dispostos a esperar.

O algoritmo proposto foi avaliado em um conjunto de dados do TRECVID 2007 (BBC Rushes) e também em vídeos do Open Video e do YouTube, e comparado com abordagens recentes na literatura de sumarização de sequências de vídeo. Os experimentos foram cuidadosamente projetados a fim de assegurar significância estatística. Resultados de uma avaliação subjetiva com usuários mostram que esse método produz resumos de vídeo com alta qualidade e velocidade computacional.

7. Conclusões e Perspectivas

Nesta tese, foi realizado um estudo abrangente em sistemas de gestão de vídeos por conteúdo, abordando temas que englobam desde a caracterização de vídeos até aspectos de indexação e armazenamento desses dados. Mais especificamente, esta tese introduziu cinco abordagens originais voltadas à análise, indexação e recuperação de vídeos digitais. O resultado final da combinação desses métodos é a especificação e implementação de

[1] `http://www.recod.ic.unicamp.br/~jurandy/vison/`

um protótipo de ambiente computacional para a gestão de vídeos que é, ao mesmo tempo, eficiente e eficaz, mantendo um padrão de qualidade comparável ao estado da arte.

Referências

[Almeida 2011] Almeida, J. (2011). *Recuperação de Vídeos Comprimidos por Conteúdo*. PhD thesis, Instituto de Computação, Unicamp, Campinas, SP, Brazil.

[Almeida et al. 2011a] Almeida, J., Leite, N. J., and Torres, R. S. (2011a). Comparison of video sequences with histograms of motion patterns. In *IEEE Int. Conf. Image Process. (ICIP'11)*, pages 3673–3676.

[Almeida et al. 2011b] Almeida, J., Leite, N. J., and Torres, R. S. (2011b). Rapid cut detection on compressed video. In *Iberoamerican Congress Pattern Recog. (CIARP'11)*, pages 71–78.

[Almeida et al. 2012a] Almeida, J., Leite, N. J., and Torres, R. S. (2012a). Online video summarization on compressed domain. *J. Visual Communication and Image Representation (JVCIR)*. DOI: 10.1016/j.jvcir.2012.01.009.

[Almeida et al. 2012b] Almeida, J., Leite, N. J., and Torres, R. S. (2012b). VISON: VIdeo Summarization for ONline applications. *Pattern Recog. Letters (PRL)*, 33(4):397–409.

[Almeida et al. 2009] Almeida, J., Minetto, R., Almeida, T. A., Torres, R. S., and Leite, N. J. (2009). Robust estimation of camera motion using optical flow models. In *Int. Symp. Visual Computing (ISVC'09)*, pages 435–446.

[Almeida et al. 2010a] Almeida, J., Minetto, R., Almeida, T. A., Torres, R. S., and Leite, N. J. (2010a). Estimation of camera parameters in video sequences with a large amount of scene motion. In *Int. Work. Systems, Signals and Image Process. (IWSSIP'10)*, pages 348–351.

[Almeida et al. 2010b] Almeida, J., Torres, R. S., and Leite, N. J. (2010b). BP-tree: An efficient index for similarity search in high-dimensional metric spaces. In *ACM Int. Conf. Inf. Knowl. Management (CIKM'10)*, pages 1365–1368.

[Almeida et al. 2010c] Almeida, J., Torres, R. S., and Leite, N. J. (2010c). Rapid video summarization on compressed video. In *IEEE Int. Symp. Multimedia (ISM'10)*, pages 113–120.

[Almeida et al. 2010d] Almeida, J., Valle, E., Torres, R. S., and Leite, N. J. (2010d). DAHC-tree: An effective index for approximate search in high-dimensional metric spaces. *J. Information and Data Management (JIDM)*, 1(3):375–390.

[Kozievitch et al. 2011a] Kozievitch, N. P., Almeida, J., Torres, R. S., Leite, N. J., Gonçalves, M. A., Murthy, U., and Fox, E. A. (2011a). Towards a formal theory for complex objects and content-based image retrieval. *J. Information and Data Management (JIDM)*, 2(3):321–336.

[Kozievitch et al. 2011b] Kozievitch, N. P., Almeida, J., Torres, R. S., Santanchè, A., and Leite, N. J. (2011b). Reusing a compound-based infrastructure for searching video stories. In *IEEE Int. Conf. Inf. Reuse Integration (IRI'11)*, pages 222–227.

[Li et al. 2011] Li, L. T., Almeida, J., and Torres, R. S. (2011). RECOD working notes for placing task mediaeval 2011. In *Working Notes Proceedings of the MediaEval 2011 Workshop (MEDIAEVAL'11)*.

[Pinto-Cáceres et al. 2011] Pinto-Cáceres, S. M., Almeida, J., Neris, V. P. A., Baranauskas, M. C. C., Leite, N. J., and Torres, R. S. (2011). Navigating through video stories using clustering sets. *Int. J. Multimedia Data Eng. Management (IJMDEM)*, 2(3):1–20.