Uma comparação entre métodos de segmentação automática de tomadas em vídeos

Welington de Jesus Lisboa [*], Carlos A. F. Pimentel Filho[**]

Celso A. Saibel Santos[#]

Resumo. *Este artigo tem como foco central a segmentação temporal de vídeos digitais, com base na detecção automática de corte de tomadas. No artigo, os resultados obtidos com a aplicação de alguns métodos apresentados na literatura foram comparados com relação à taxa de acerto na detecção das tomadas ou segmentos de vídeo obtidos.*

Palavras chave: Segmentação temporal de vídeo, detecção de corte de tomadas, segmentação automática de vídeos.

1. Introdução

Aplicações de vídeo e imagens digitais são importantes para educação, entretenimento, e muitas outras aplicações multimídia. Com o rápido crescimento de equipamentos de captura de vídeo de baixo custo, tais como *webcams*, celulares com recurso de gravação de vídeo, e mesmo filmadoras tradicionais, observamos um crescimento no volume de arquivos de vídeo digitais.

Novas aplicações em vídeo digital permitem outras estruturas de navegação e busca de conteúdo mais avançadas que as conhecidas do sistema VHS, por exemplo. Portanto é possível usar uma nova abordagem de navegação em vídeos com a tecnologia digital. A delimitação de tomadas (ou *shots*) de vídeo é a base para a

---

[*] Aluno do 2º ano de graduação em Ciência da Computação da UNIFACS.

[**] Aluno do Mestrado Acadêmico em Sistemas e Computação da UNIFACS.

[#] Orientador. Professor Titular da UNIFACS.

estruturação de vídeo, agregando quadros contíguos em seqüência com o mesmo contexto (Tekalp, 2000).

Além de possibilitar a navegação do vídeo entre cortes de tomadas, a detecção de corte das mesmas é importante em aplicações de extração automática de características, por exemplo: o tamanho médio dos *shots***,** em *trailers* de filmes, é uma característica usada, dentre outras, para classificar filmes em um dos quatro gêneros: comédia, drama, terror ou ação (Rasheed Z et al, 2002).

Neste artigo são apresentados conceitos básicos sobre a estrutura de vídeos digitais, resolução espacial e temporal de vídeo, espaços de cores e delimitações de tomadas. São mostrados também, três métodos de segmentação de tomadas, os resultados práticos obtidos e a comparação do desempenho dos métodos com o resultado esperado.

2. Vídeos digitais e segmentação temporal

Vídeo é uma seqüência de imagens ou fotogramas que, quando reproduzidas uma a uma, em determinada velocidade, apresentam a ilusão de movimento. Tipicamente são usados 30 fps[1] para se obter tal efeito. Um vídeo pode ainda conter um canal de áudio sincronizado.

Existem muitas terminologias usadas para descrever atributos de um vídeo. Nessa seção, traduzimos e adaptamos a terminologia de (Smith & Chen, 2000) de atributos e padrões em vídeo e imagem, mostrada na tabela 1.

A segmentação temporal de vídeo refere-se à identificação de quadros do vídeo com algum grau de homogeneidade (Tekalp, 2000). Consideramos que um segmento de vídeo consiste em dois ou mais conjuntos de quadros em um vídeo. Em outras palavras, a segmentação temporal vai identificar intervalos de tempo no vídeo onde ocorrem mudanças rápidas do que estava sendo filmado, independente do seu contexto, ou simplesmente, um corte de câmera.

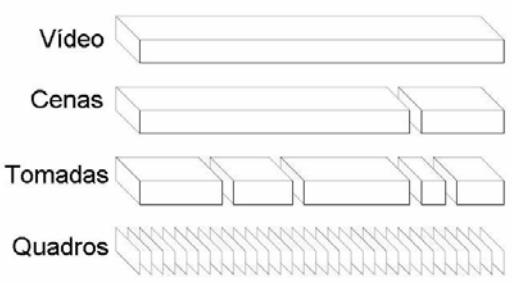
Os autores receberam apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq através do projeto DEIVID: Ambiente para DEscrição, Indexação e Consulta de Conteúdos de VÍdeos Digitais (Ref 506647/2004-8)

Tabela 1 – Terminologia de vídeo

Termo	Comentário
Vídeo (Imagem/Áudio)	O termo vídeo será usado para representar um fluxo de imagens e áudio.
Cena (Imagem)	Uma cena é uma unidade de imagens que carregam alguma semântica.
Quadro ou *Frame*	Refere-se a um único fotograma do vídeo
Segmento	Um subconjunto homogêneo de quadros do vídeo, delimitados no tempo por semântica ou não, embora o presente trabalho não considere a semântica.
Tomada	Conjunto de *Frames* consecutivos entre um corte de câmera e outro.
Áudio	Refere-se ao canal de áudio associado ao vídeo

Uma tomada consiste de um ou mais quadros cujas características são semelhantes. Estes quadros são gerados e gravados de forma contígua, representando uma ação em relação ao tempo e espaço. As cenas são uma combinação de várias tomadas num mesmo contexto, por exemplo, um diálogo entre personagens. Por sua vez, um conjunto de cenas compõe um vídeo, como exibido na Figura 1 (Santos T., 2004).

A aquisição automática de corte de tomadas de vídeos tem várias aplicações. Uma das mais importantes é a determinação de quadros chave, que carregam informação visual da tomada como um todo, podendo funcionar como uma espécie de resumo. Quadros chave também são usados para a construção de mosaicos (IBM MArvel, 2006). Os mosaicos representam imagens construídas de vários quadros chaves pequenos, assim, os mesmos, gerados a partir dos quadros chaves, podem representar em uma única imagem a idéia de toda a ação contida em um segmento do vídeo. (Santos T., 2004). Um mosaico é mostrado na figura 2.



[1] O termo fps indica *frames per second* ou quadros por segundo.

Figura 1: Estruturação de Vídeo. Fonte: (Santos T., 2004)

Quadros-chaves e mosaicos podem ser utilizados na criação de índices, através da extração de características de imagem, como cor, textura e forma. Por sua vez, as tomadas são adequadas para extração de características envolvendo movimento.

Figura 2 – Mosaico de Vídeo. Fonte: (IBM MArvel, 2006)

3 - Métodos de segmentação de tomadas

Nesse tópico apresentamos os métodos de detecção automática de corte de tomadas analisados. De modo geral, os métodos de detecção automático de corte de tomadas baseiam-se na busca de similaridade entre quadros sucessivos, classificando automaticamente pontos de baixa similaridade.

Neste artigo três métodos para segmentação de tomadas são avaliados: Comparação *pixel-a-pixel* (Santos T., 2004), comparação de distância dos histogramas RGB (Smith & Chen, 2000) e mínimo dos histogramas HSV (Rasheed Z. et al, 2002).

3.1 - Comparação pixel-a-pixel

Uma das técnicas mais simples e de fácil implementação para a detecção de tomadas de vídeo é através da soma das diferenças (d_p) entre os *pixels* de quadros adjacentes na mesma posição espacial. Esse método produz em valores baixos de d_p para imagens semelhantes e valores mais altos ou “picos” para imagens diferentes. Tal técnica depende da fixação de um limiar k_p, que é definido experimentalmente. Os valores que ultrapassarem esse limiar, podem ser então considerados como um corte de tomada. Para fixarmos um limiar independente da resolução espacial dos quadros, dividimos o somatório pela quantidade total de pixels do quadro. A equação para o cálculo da comparação *pixel-a-pixel* é definida como se segue:

$$d_p(q_i, q_j) = \frac{1}{MN} \sum_{x=1}^{M} \sum_{y=1}^{N} | q_i(x,y) - q_j(x,y) | \qquad \text{eq.1}$$

Na eq.1, q_i (x,y) é a intensidade do *pixel* (*x*,*y*) que pode indicar, para quadros monocromáticos, um nível de cinza (q_i de resolução *M*x*N*), ou, para quadros coloridos, uma valor no domínio RGB (*red, green, blue*).

$$d_p(q_i, q_j) = \frac{1}{3MN} \sum_{c \in \{R,G,B\}} \sum_{x=1}^{M} \sum_{y=1}^{N} | q_i(x,y,c) - q_j(x,y,c) | \qquad \text{eq.2}$$

Na eq.2, um quadro q_i delimita uma fronteira de tomada se $d_p(q_i,q_{i+1}) > k_p$, onde k_p é um limiar definido experimentalmente.

3.2 - Comparação de Histogramas RGB

A diferença de histogramas é menos sensível a movimentos súbitos de objetos na cena, sendo uma forma bastante efetiva de detecção de similaridade de imagens.

$$d_h(q_i, q_j) = \sum_{l=1}^{L} | h_i[l] - h_j[l] | \qquad \text{eq.3}$$

Dizemos que um quadro q_i delimita uma tomada se $d_h(q_i,q_{i+1}) > k_h$, onde k_h é um limiar definido experimentalmente.

3.3 – Somatório dos mínimos dos histogramas HSV:

Diferentemente do RGB, no modelo de cor HSV a representação das cores em uma imagem é feita através das componentes *Hue* (croma), *Saturation* (saturação) e *Value* (valor). Com estas variáveis, o modelo HSV aproxima-se muito do modelo intuitivo utilizado em artes visuais, onde os conceitos qualitativos de matiz, luz e tonalidade são mais empregados.

O método consiste em somar o menor valor de cada histograma HSV das imagens em comparação. A figura 3 mostra um exemplo de imagem representada no formato HSV (Rasheed Z et al, 2002).

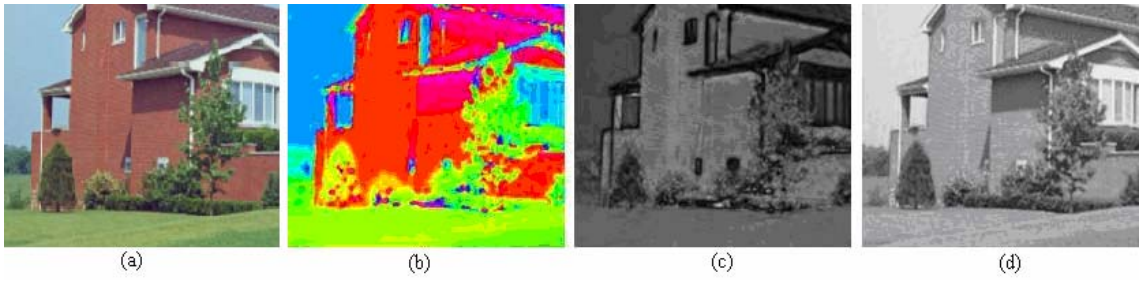


Figura 3: (a) imagem original, (b) Croma (*H*), (c) Saturação (*S*) e (d) Valor (*V*).

Sendo *S*(*i*) função dos histogramas H_i e H_{i-1} os quadros *i* e *i*-1 respectivamente:

$$S(i) = \sum_{j \in allbins} \min(H_i(j), H_{i-1}(j)) \qquad \text{eq.4}$$

Como resultado, obtemos um sinal *S*(*i*) com muito ruído, que é suavizado usando o método iterativo de "*Gaussian Kernel*" (Rasheed Z et al, 2002). Isso porque a variância da função de Gauss varia de acordo com o sinal do gradiente. Formalmente:

$$S^{t+1}(i) = S^t(i) + \lambda\left[c_E . \nabla_E S^t(i) + c_W . \nabla_W S^t(i)\right] \qquad \text{eq.5}$$

Onde: *t* é o número da iteração; $0 < \lambda < 1/4$ e:

$$\nabla_E S(i) \equiv S(i+1) - S(i), \qquad \text{eq.6}$$

$$\nabla_W S(i) \equiv S(i-1) - S(i). \qquad \text{eq.7}$$

Os coeficientes condicionais são funções dos gradientes atualizados a cada iteração:

$$c_E^t = g\left(\left|\nabla_E S^t(i)\right|\right) \qquad \text{eq.8}$$

$$c_W^t = g\left(\left|\nabla_W S^t(i)\right|\right) \qquad \text{eq.9}$$

Onde: $\mathbf{g}(\nabla S) = e^{-\left(\frac{|\nabla E|}{K}\right)^2}$

De acordo com (Rasheed Z et al, 2002a)) foram usados os seguintes valores para $\lambda = 0.1$ e $k = 0.1$. Finalmente, as fronteiras de tomadas são indicadas pelos mínimos locais da função *S*(*i*), pois nesse método não há um limiar a ser fixado.

4 - Análise e resultados

Os experimentos envolveram a aplicação dos métodos selecionados em um conjunto de 4 vídeos, com diferentes características em termos de duração, tipo de conteúdo, formato de codificação do vídeo. Os algoritmos de detecção foram executados numa arquitetura Pentium IV 3GHz, 1GB de RAM, especialmente dedicada ao processamento. Os processamentos levaram cerca de 3 vezes o tempo total do vídeo em todos os casos. Os resultados obtidos nos experimentos são apresentados na tabela 2.

Tabela 2 – Relação de cortes detectados nos três métodos.

Nome do vídeo	Man	PP	RGB	HSV
Greenday – Minority	128	119	121	110
Creed – Higher	66	220	113	91
Leonardo de Caprio vs Jack Nicholson	21	49	18	33
Star Wars vs. Alien	75	60	87	52

Legenda: Man = Segmentação Manual, PP = Comparação pixel-a-pixel, RGB = Diferença dos Histogramas RGB e HSV = Mínimos dos histogramas HSV.

A partir dos resultados na tabela 2, foi possível verificar que nenhum dos métodos apresentou desempenho superior aos demais para todos os casos analisados. Entretanto, na maior parte dos casos, o método de segmentação baseado na diferença dos histogramas RGB foi o que apresentou resultados mais próximos do que poderia ser obtido com a segmentação manual dos vídeos. A tabela 3 apresenta os resultados da aplicação dos métodos com a taxa média de erro, e detecção de falsos positivos.

Tabela 3 – Taxa de erros e detecção de falsos positivos.

Método	Taxa de média de erro	Falsos positivos
PP	84,9%	283%
RGB	16,8%	143%
HSV	12,5%	147%

5. Conclusão

Este artigo apresentou um estudo sobre três métodos de segmentação automática de tomadas de vídeo. Os métodos foram empregados em diferentes tipos de vídeo e com diferentes durações.

O método de detecção *pixel-a-pixel* mostrou-se sensível ao movimento de objetos no quadro, isto se torna uma desvantagem, já que se houver movimento é possível encontrar falsos cortes de tomada ou falsos positivos. Isso ocorre, por exemplo, quando existem efeitos de zoom no vídeo, para os quais apenas a mudança no tamanho dos objetos tem a tendência de gerar falsos positivos.

O método da diferença dos histogramas RGB tem que contornar o problema de que quadros completamente diferentes possam ter histogramas muito similares, que podem também gerar falsos positivos. Entretanto, a probabilidade destas situações ocorrerem é baixa, comprometendo pouco a eficácia do método. Esses casos ocorrem principalmente em cortes de tomadas no mesmo ambiente, que por conseguinte, tendem a apresentar a mesma taxa de iluminação. Por outro lado, a diferença de histogramas apresenta resistência a movimentação de objetos, uma vantagem sobre o método de comparação *pixel-a-pixel*.

Esperava-se, de acordo com a literatura (Rasheed Z et al, 2002), que o método do somatório dos Histogramas HSV, apresentasse os melhores resultados para a detecção de tomadas. Entretanto, não foi exatamente o que aconteceu nos experimentos realizados. O mau desempenho apresentado, entretanto, não determina que o método é sempre ineficiente. Isso porque, o método tem algumas variáveis a serem ajustadas de forma heurística, como λ, k e a quantidade de iterações da função de suavização. Um outro problema do método é a detecção de mínimos locais. Como o método não trabalha com um limiar, é preciso descobrir os mínimos locais, que são realmente relevantes para indicação de delimitação de tomadas de vídeo.

Referências

(IBM MArvel, 2006) Site: <http://mp7.watson.ibm.com/marvel/> Ultimo acesso: 29 de Agosto de 2006

(Rasheed Z et al, 2002) Zeeshan Rasheed, Yaser Sheikh e Mubarak Shah – Semantic Film Preview Using Low-Level Computable Features - 2002.

(Santos T., 2004) Tiago Texeira Santos - Segmentação automática de tomadas em vídeo Dissertação de Mestrado - Universidade de São Paulo - 2004.

(Smith & Chen, 2000) Micheal A. Smith & Tsuhan Chen - Image and Video Indexing and Retrievel - Hand Book of image and video processing - Editor Al Bovik 2000

(Tekalp, 2000) A. Murat Tekalp - Video Segmentation - Hand Book of image and video processing – Editor Al Bovik 2000