Detecção de Algumas Transições Abruptas em Seqüências de Imagens

Nielsen Cassiano Simões

Dissertação de Mestrado

Instituto de Computação
Universidade Estadual de Campinas

Detecção de Algumas Transições Abruptas em Seqüências de Imagens

Nielsen Cassiano Simões[1]

31 de maio de 2004

Banca Examinadora:

- Prof. Dr. Neucimar Jerônimo Leite (Orientador)
- Prof. Dr. Silvio Jamil Ferzoli Guimarães
  Pontifícia Universidade Católica de Minas Gerais
- Prof. Dr. Alexandre Xavier Falcão
  Universidade Estadual de Campinas
- Prof. Dr. Siome Klein Goldenstein (Suplente)
  Universidade Estadual de Campinas

[1]**Suporte financeiro da CAPES e apoio parcial do SAI/PRONEX-MCT.**

Detecção de Algumas Transições Abruptas em Seqüências de Imagens

Este exemplar corresponde à redação final da Dissertação devidamente corrigida e defendida por Nielsen Cassiano Simões e aprovada pela Banca Examinadora.

Campinas, 26 de fevereiro de 2004.

Prof. Dr. Neucimar Jerônimo Leite
(Orientador)

Dissertação apresentada ao Instituto de Computação, UNICAMP, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

"De tudo, ficaram três coisas: a certeza de que ele estava sempre começando, a certeza de que era preciso continuar e a certeza de que seria interrompido antes de terminar. Fazer da interrupção um caminho novo. Fazer da queda um passo de dança, do medo uma escada, do sonho uma ponte, da procura um encontro."

(*Fernando Sabino*)

Resumo

Um vídeo digital é composto por uma seqüência de imagens ou quadros. Uma tomada de vídeo corresponde a um subconjunto de quadros associados a um segmento ininterrupto no tempo, espaço e configurações gráficas de tela. O problema de detecção de transições entre tomadas é o primeiro e um dos mais importantes passos para o processo de análise e segmentação estrutural de um vídeo digital. Para a detecção automática destes eventos, muitos trabalhos consideram métricas de dissimilaridades entre quadros baseadas, por exemplo, em informações de cor, forma e textura, enquanto outros aplicam técnicas de processamento de imagens sobre uma imagem representativa do vídeo. Esta dissertação apresenta uma nova abordagem para detecção de transições e efeitos abruptos (cortes e flashes) em seqüências de imagens, utilizando algoritmos simples e de baixo custo computacional, baseados na identificação de padrões em sinais $1\mathbb{D}$. Os resultados apresentados ilustram o bom desempenho dos métodos na identificação dos eventos considerados.

Abstract

A digital video is represented by a sequence of images or frames. A video shot is an uninterrupted segment of screen time, space and graphical configurations. The problem of transition detection between shots can be seen as one of the most important step to the process of segmenting and parsing a digital video. In order to have an automatic detection of these events, some researches consider frame by frame comparison using dissimilarity measures based on color, form and texture information, while others apply images processing techniques over a representative image of the whole video. This work describes a new approach to detect transitions and abrupt effects (cuts and flashes) in image sequences, by considering simple and low computational cost algorithms which are defined based on patterns identification of a $1\mathbb{D}$ signal. The results presented here show the good performance of the method in the identification of the corresponding events.

Agradecimentos

Agradeço...

A Deus, em primeiro lugar, pela inspiração, força e sustento que tem me condedido para a conclusão dessa dissertação.

Ao meu orientador, Neucimar, sem o qual este trabalho não teria sido possível. Por seu apoio, confiança, sua paciência, e por todo o meu aprendizado e meu crescimento acadêmico, obtidos todos esses anos.

À minha esposa, Priscila, pelo amor, carinho, incentivo e compreensão dedicados nesse período.

Aos meus pais, Pedro e Mirtes, e a meu irmão, Nemer, por acreditarem na minha educação.

Aos demais familiares pela confiança em meu trabalho.

À profa. Cláudia, pelas dicas, críticas e contribuições, e pelas reuniões do grupo LIS; e a todos que por lá passaram nestes últimos três anos.

Aos meus amigos, companheiros de turma e de corredor, pela troca de conhecimentos e experiências.

A todos os funcionários do IC, pela contribuição direta e indireta.

À CAPES e ao projeto SAI-PRONEX/MCT pelo apoio e suporte financeiro, e por acreditarem na pesquisa acadêmica.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Até recentemente, os vídeos analógicos ofereciam uma qualidade regular de som e imagem, e consumiam altas horas de trabalho de edição para que sua versão final, seja esta um filme, um comercial ou um documentário, por exemplo, ficasse em condições de satisfazer os expectadores exigentes. Atualmente, os crescentes avanços da tecnologia, principalmente na área computacional, nos levaram à definição do vídeo digital.

O vídeo digital é hoje muito utilizado, principalmente, por oferecer maiores facilidades de edição, qualidade de gravação e outros recursos disponíveis em um computador. Atualmente, a facilidade em se criar e editar vídeos, o crescente aumento da popularização das câmeras digitais e a utilização em larga escala das *web cams*, têm tornado o vídeo digital cada vez mais popular. Outro ponto importante está na facilidade em se compartilhar esses vídeos, além da crescente utilização dos vídeos por demanda nos principais portais da Internet.

O avanço nas tecnologias de compressão de dados para a transmissão de vídeo, na velocidade de acesso da Internet e na capacidade de armazenamento a um custo mais reduzido, têm dado sua contribuição neste sentido. Não é difícil encontrar, nos programas de compartilhamento de arquivos pela Internet, os famosos *Pear to Pear (P2P)*, vídeos inteiros de clipes, filmes, séries de TV, vídeos caseiros etc.

Todo esse crescimento na utilização de vídeo digital leva à necessidade da criação de bibliotecas virtuais, que já vêm sendo estudadas nos dias de hoje, inclusive com protótipos disponíveis na internet tais como, o *Infomedia Digital Video Understanding* [WCHKdu], *VISION Digital Video Library System* [GGml] e o *VideoQ: An objected oriented video search engine* [CHoQ]. Várias dessas bibliotecas digitais possuem em seus acervos horas de vídeos digitais armazenados, disponibilizados muitas vezes para diversos tipos de usuários. Várias pesquisas procuram realizar eficientemente o armazenamento e a recuperação de vídeos digitais em grandes bancos de dados.

Além disso, com o avanço nas pesquisas relacionadas à busca por conteúdo em banco

de dados de imagens, inúmeros trabalhos começaram a surgir ampliando essa busca por conteúdo para banco de dados de vídeo [CL97]. Todos esses problemas envolvem, principalmente, a necessidade de indexação de vídeos em bancos de dados, ou a decomposição do vídeo em fragmentos. Para tanto, a estrutura do vídeo e, se possível, a sua semântica devem ser extraídos e representados eficientemente. Uma das primeiras etapas desta segmentação consiste da detecção das tomadas da seqüência de vídeo digital.

A seguir, serão apresentados e definidos alguns dos conceitos básicos utilizados no decorrer desta dissertação.

1.1 Conceitos Básicos

Nesta seção, são apresentados os principais conceitos empregados no restante desta dissertação relacionados, principalmente, com modelo ou estrutura de um vídeo digital.

1.1.1 Vídeo Digital: Algumas Definições

Um vídeo digital, ou um segmento de vídeo digital, pode ser subdividido nos seguintes itens, conforme ilustra a Figura 1.1:

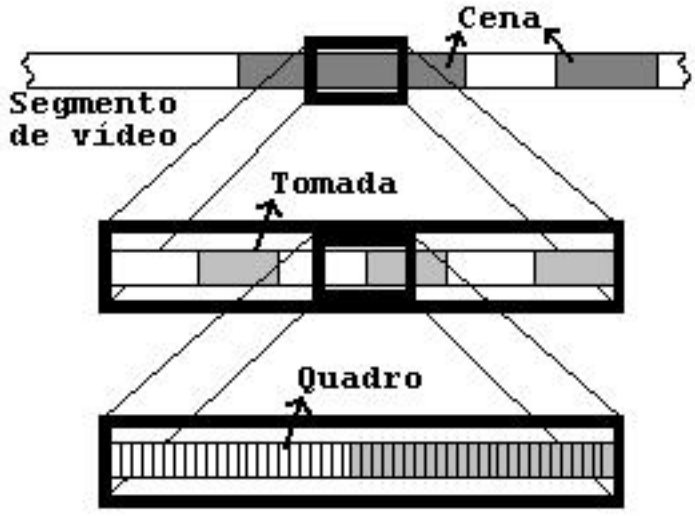


Figura 1.1: Divisão lógica e física de um segmento de vídeo digital.

Definição 1.1. Quadro ou Frame - *Um quadro é uma foto instantânea do vídeo digital. É uma imagem referente a um determinado instante captado por uma câmera. Pode ser visto como a menor parte da estrtutura de um vídeo digital. Para um quadro f, de altura H e largura W, $f(x, y)$ representa a intensidade do pixel (x, y) do quadro f, em que $x \in [0, W-1]$ e $y \in [0, H-1]$.*

Definição 1.2. Video - *Um vídeo é uma coletânea de imagens agrupadas seqüencialmente conforme uma relação temporal. Seja V um segmento de vídeo digital com N quadros de largura W e altura H. Um vídeo digital pode ser definido como:*

$$V = (f_0, f_1, \ldots, f_{N-1}),$$

em que f_t representa o t-ésimo quadro do segmento de vídeo digital.

Definição 1.3. Tomada - *Uma tomada é uma sequência do vídeo digital sem interrupção das configurações espaciais, gráficas ou temporal. Em geral, uma tomada representa um intervalo de tempo contínuo capturado por uma câmera durante uma filmagem. Seja S uma tomada com K quadros no segmento de vídeo, iniciada a partir do quadro f_j, $\forall j \in [0, N-K]$. A tomada S pode então ser definida como:*

$$S = (f_j, \ldots, f_{j+K-1})$$

Definição 1.4. Cena - *Uma Cena é um conjunto de tomadas agrupadas segundo uma característica semântica do tempo, espaço ou gráfico. As cenas estão relacionadas diretamente com o conteúdo de parte do segmento de vídeo digital, compostas de uma ou mais tomadas.*

Definição 1.5. Quadro-chave - *Um quadro-chave (key-frame) é um quadro capaz de representar o melhor possível uma unidade lógica do segmento de vídeo digital. Em geral, um quadro-chave representa uma tomada mas também pode representar uma cena. Quadros-chaves são utilizados, por exemplo, na compressão e indexação de um vídeo digital.*

Cada tipo de vídeo possui uma característica própria, dependendo do objetivo proposto pelo realizador. Em geral, quando cenas são particionadas e exibidas entre outras cenas transmite-se a idéia, em termos semânticos, de diversos acontecimentos simultâneos.

1.1.2 Transições

Normalmente, um segmento de vídeo, após sua filmagem, é submetido a um processo de edição. É nessa etapa que surgem os principais efeitos de transições entre tomadas.

Definição 1.6. Transição - *Uma transição pode ser definida como qualquer interrupção no segmento de vídeo referente a uma tomada, ocasionada pelo processo de filmagem ou de edição do vídeo digital.*

As transições são normalmente efeitos decorrentes da edição de um vídeo, que permitem ao editor dar uma maior dinâmica ao mesmo, realçando o conteúdo do vídeo. Em um vídeo com cenas de ação, por exemplo, é muito comum se ter um número elevado de cortes e tomadas curtas. Normalmente, em um documentário, as tomadas são maiores e muitos cortes são utilizados. Em filmes de suspense, as tomadas tendem a ser mais longas, separadas principalmente por cortes, com intervalos de duração decrescente conforme se aumenta a tensão dos fatos. Serão descritos abaixo os principais efeitos de transição, classificados segundo sua apresentação visual como **Transições Abruptas** e **Transições Graduais**.

Transições Abruptas

As transições abruptas são transições decorrentes da junção de duas tomadas, com duração nula, i.e., nenhum quadro é criado ou modificado durante a transição.

Definição 1.7. Corte - *Um corte é uma transição de tempo nulo onde duas tomadas são anexadas uma ao final da outra, sem a criação de novos quadros ou modificações dos quadros existentes (Figura 1.2).*

Figura 1.2: Exemplo de um corte em um segmento de vídeo digital.

Transições Graduais

As transições graduais são utilizadas, geralmente, em cenas com menos ação ou para interrupção ou término de algum acontecimento que continua em outro, relativamente desconexo ou distante [Chaml], e geralmente aparecem na mudança entre cenas. Estas transições envolvem a criação ou alteração de alguns quadros durante a transição, normalmente estes quadros são uma mistura de dois quadros de tomadas diferentes, variando a taxa dos pixels de cada um dos quadros presentes na transição. *Fades*, *Dissolves*, *Wipes* e *Morphings*, discutidos a seguir, são exemplos de transições graduais.

Definição 1.8. Fade-In - *Um fade-in é uma transição de um quadro monocromático (preto ou branco), normalmente preto, para o quadro da tomada a ser iniciada. Seja S_t o t-ésimo quadro de uma tomada S. Os frames de um fade-in de T quadros, a partir de um quadro P (preto ou branco), para a tomada S, podem ser obtidos por:*

$$Fd_t = \alpha(t) \times P + (1 - \alpha(t)) \times S_t,$$

em que $t \in [0, T-1]$, e $\alpha(t)$ é uma função de transformação, normalmente linear decrescente.

Figura 1.3: Exemplo de um fade-in em um segmento de vídeo digital.

Geralmente, o fade-in é utilizado no início de um vídeo digital, mas pode aparecer também no decorrer do mesmo.

Definição 1.9. Fade-Out - *O fade-out é uma transição de um quadro de uma tomada qualquer para um quadro monocromático (preto ou branco), normalmente preto. Seja S_t o t-ésimo quadro de uma tomada S. Os frames de um fade-out de T quadros, a partir de um quadro P (preto ou branco), para a tomada S, podem ser obtidos por:*

$$Fo_t = (1 - \alpha(t)) \times P + \alpha(t) \times S_t,$$

em que $t \in [0, T-1]$, e $\alpha(t)$ é uma função de transformação, normalmente linear decrescente.

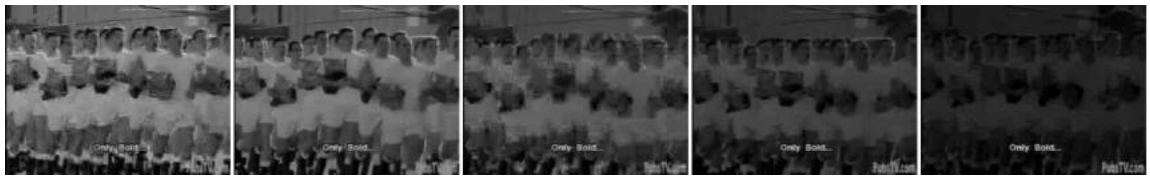


Figura 1.4: Exemplo de um fade-out em um segmento de vídeo digital.

O fade-out é normalmente utilizado no final de um vídeo digital, mas também pode aparecer no decorrer do mesmo.

Definição 1.10. Dissolve - *Um dissolve é uma transição em que dois quadros, um de cada tomada, são misturados de forma que a taxa de pixels de alguns quadros da última tomada se reduz na mesma proporção em que a taxa de pixels dos quadros da próxima tomada aumenta. Sejam S_{1_t} e S_{2_t} o t-ésimo quadro das tomadas S_1 e S_2, respectivamente. Os frames do dissolve de T quadros da tomada S_1, de K quadros, para a tomada S_2 podem ser obtidos por:*

$$Ds_t = (1 - \alpha(t)) \times S_{1_{K-1-t}} + \alpha(i) \times S_{2_t}$$

em que $K \geq T$, $t \in [0, T-1]$, e $\alpha(t)$ é uma função de transformação, normalmente linear.

Figura 1.5: Exemplo de um dissolve em um segmento de vídeo digital.

O Dissolve caracteriza uma transição mais lenta entre duas tomadas, reduzindo o impacto visual da transição em um vídeo.

Definição 1.11. Wipe - *Um wipe é um grupo de transições onde se pode observar uma ou mais linhas imaginárias separando os quadros da última tomada dos quadros da nova tomada.*

Existem vários tipos de padrões de wipes e alguns destes padrões estão ilustrados na Figura 1.6, obtidos em [Joyns].

Definição 1.12. Morphing - *Um morphing é uma transição gradual derivada de uma função matemática especial, entre os pixels dos quadros de duas tomadas consecutivas, com o objetivo de minimizar a percepção da alteração entre as tomadas. O morphing também pode ser encontrado como um efeito especial para transições de objetos internos de uma tomada.*

Devido a sua complexidade para criação, a utilização do morphing como efeito é maior que como transição entre tomadas, e nenhum dos trabalhos relacionados nessa dissertação consideram este evento para transições entre tomadas.

As transições graduais acontecem com menor freqüência nos principais vídeos e são, na maioria das vezes, mais difíceis de serem detectadas automaticamente. Geralmente, os cortes representam 95% das transições em filmes mas as atuais produções de programas de televisão tendem a reduzir esta taxa para, no máximo, 60% do total das transições [AM88].

Figura 1.6: Exemplos de alguns padrões de wipes em um segmento de vídeo digital: a) *diagonal*, b) *door open*, c) *iris round*, d) *push vertical* e e) *split*.

1.1.3 Trabalho de Câmera

Durante o processo de filmagem, alguns efeitos de câmera ou trabalhos de câmera (*camera works*), ou ainda operações de câmera, podem ser utilizados. Esses efeitos podem ser obtidos a partir da movimentação da filmadora sob uma determinada direção, pela variação do zoom (*zoom-in* ou *zoom-out*) ou ainda pela variação do ângulo de filmagem, realizados pelo operador. A Figura 1.7 ilustra alguns dos principais trabalhos de câmera.

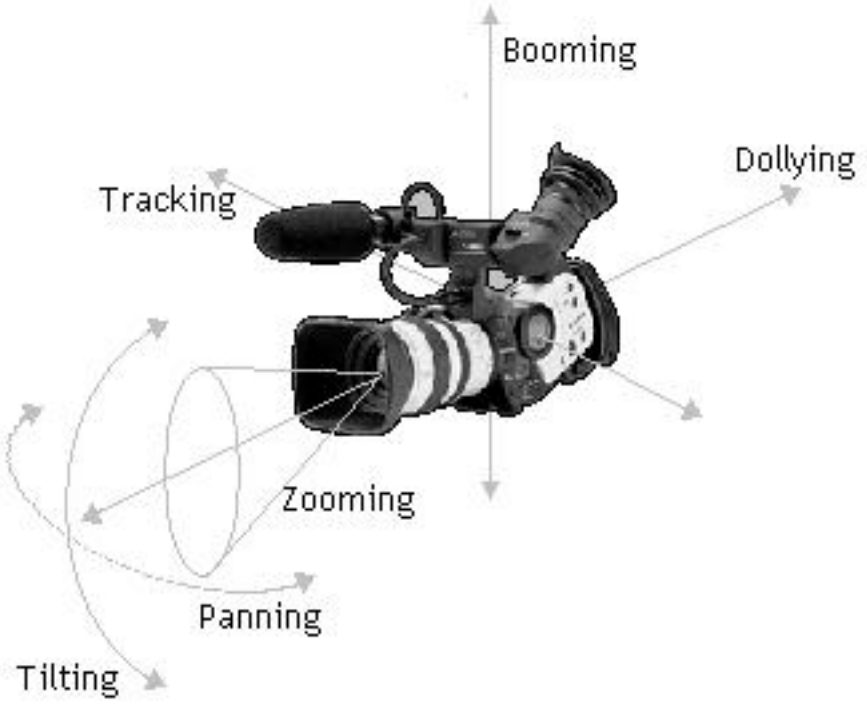


Figura 1.7: Ilustração dos principais trabalhos de câmera utilizados.

Além dos trabalhos de câmera apresentados nesta figura, o *flash* é outro efeito freqüentemente presente em vídeos de telejornais. Este se caracteriza pela presença de alta luminosidade durante poucos quadros, normalmente decorrentes de flash de câmeras de fotografias. Durante a filmagem, esse acontecimento está, em geral, associado à presença de pessoas importantes e famosas. A Figura 1.8 ilustra uma ocorrência de flash em um segmento de vídeo.

Figura 1.8: Exemplo da ocorrência de um flash em um segmento de vídeo digital.

1.2 Proposta e estrutura da dissertação

Essa dissertação propõe um novo método para detecção de tomadas em segmentos de vídeo digital, simples e eficiente que primeiramente realiza uma transformação do vídeo para uma imagem representativa e, posteriormente, uma outra transformação para um vetor representativo, convertendo o problema de detecção de transições em vídeo para um problema de identificação de padrões em sinais $1\mathbb{D}$. Além disso, também é proposto um algoritmo para a detecção de flashes, considerando a análise de sinais $1\mathbb{D}$ extraídos da imagem representativa e de informações dos quadros do vídeo digital.

O restante desta dissertação está organizado da seguinte maneira: o capítulo 2 expõe as principais técnicas para detecção de transições relacionadas a esta dissertação. O capítulo 3 detalha o método proposto e o trabalho realizado. O capítulo 4 ilustra alguns resultados obtidos durante este trabalho. Finalmente, o capítulo 5 contém as conclusões dessa dissertação e discute os trabalhos futuros relacionados.

Capítulo 2

Detecção de Transições

O processamento manual (sumarização, anotação, etc) de um vídeo digital permite gerar inconsistências ou erros provocados pelo operador, seja por má interpretação ou por fadiga, pois esse processo pode demandar horas para cada vídeo, dependendo do seu tamanho. A detecção e identificação das transições de tomadas, ou simplesmente detecção de tomadas, em vídeo digital, é a primeira e uma das mais importantes etapas no processo de segmentação e análise de vídeo digital. De uma forma geral, os processos envolvidos na análise de vídeo digital podem ser ilustrados conforme a Figura 2.1.

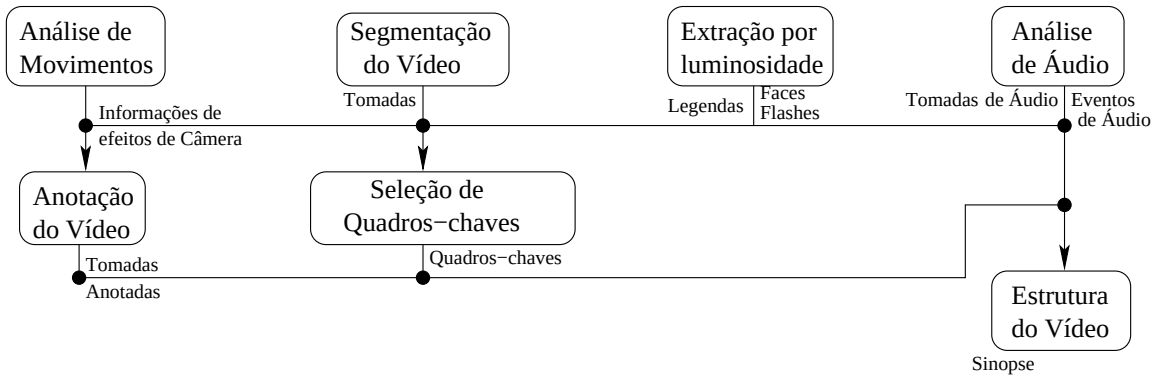


Figura 2.1: Processos envolvidos na análise de vídeo digital.

Detectar automaticamente tomadas em vídeo digital é uma tarefa fundamental na detecção de cenas, extração de quadros-chaves, e indexação. Os principais trabalhos relacionados à detecção de tomadas podem ser divididos em duas categorias:

- Aqueles aplicados diretamente ao vídeo digital, utilizando alguma **medida de dissimilaridade** definida para a detecção das tomadas.

- Aqueles aplicados sobre uma **imagem representativa** do vídeo digital, definindo um padrão característico da tomada. O método apresentado nesta dissertação enquadra-se nesta categoria.

Estes trabalhos serão relacionados em cada uma dessas categorias nas próximas seções. Algumas das técnicas apresentadas aqui são aplicadas diretamente nos quadros do vídeo digital, i.e., no domínio espacial, enquanto outras são diretamente aplicadas no domínio de compressão, mais precisamente no formato de compressão MPEG definido em [MPE95] como um padrão ISO de compressão de seqüência de imagens.

2.1 Medidas de Dissimilaridade

Devido ao grande avanço nos trabalhos relacionados à busca por conteúdo em banco de dados de imagens, muitas técnicas de indexação e medidas para esse fim foram propostas [BMM99, NPC98, RHC99, SC96]. Algumas das abordagens de segmentação de vídeo digital são baseadas em medidas de dissimilaridades derivadas do problema de indexação e recuperação de imagens [BM99, RHC99]. Em geral, as medidas de dissimilaridade são empregadas na comparação de quadros consecutivos do vídeo. Tais medidas podem estar relacionadas a informações de cores, textura, forma e movimento de objetos. Alguns desses trabalhos são descritos nas próximas seções de acordo com a característica de cada medida empregada. Para todos os casos, é assumido o intervalo $t \in [0, N-1]$, para um vídeo digital V qualquer com N quadros.

2.1.1 Informações de Cores

De forma geral, podemos subdividir as abordagens que utilizam informação de cores segundo a característica da medida de dissimilaridade escolhida. Algumas dessas técnicas são aplicadas utilizando somente a intensidade de cor dos quadros do vídeo digital, enquanto outras utilizam o histograma (Definição 2.1, a seguir) de cada quadro. Além de fornecer uma informação global dos quadros, o histograma possui a característica de ser invariante à rotação e translação.

Definição 2.1. Histograma - *O Histograma $H_{ist}(j)$ representa a quantidade de pixels de uma imagem ou de um quadro, cujo nível de cinza é igual a j, e $j \in [0, L-1]$, para L diferentes níveis de cinza.*

Diferença entre Quadros

Uma das primeiras técnicas para detecção de cortes baseia-se na diferença entre pixels de dois quadros consecutivos de um vídeo digital, conhecida como *pixel-wise comparison*.

Alguns trabalhos que utilizam esta técnica para detecção de cortes estão relacionados em [KC01, PS96, ZKS93]. Tais trabalhos consideram a diferença absoluta entre a intensidade dos pixels de dois quadros consecutivos de um vídeo digital. As Equações 2.1 e 2.2, utilizadas para imagens em níveis de cinza e coloridas, respectivamente, representam a soma do valor absoluto da diferença dos pixels entre os quadros f_t e f_{t-1}, de altura H e largura W, em que $c \in \{R, G, B\}$ corresponde à intensidade da componente de cor no espaço RGB.

$$D_t = \frac{1}{W * H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} |f_{t-1}(x, y) - f_t(x, y)| \tag{2.1}$$

$$D_t = \frac{1}{W * H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} \sum_{c} |f_{t-1}(x, y)_c - f_t(x, y)_c| \tag{2.2}$$

Dessa forma, um corte é detectado quando o valor de D_t é superior a um determinado limiar pré-especificado. O grande problema dessa abordagem é que, além do custo computacional para comparar todos os pixels de cada quadro com seu quadro imediatamente anterior, a presença de rápidos movimentos de objetos e câmera acabam sendo detectados como cortes. Ela também não permite diferenciar grandes alterações em uma região pequena, nem mesmo pequenas alterações em uma grande região do quadro. Uma abordagem apresentada em [BMM96, BMM99] utiliza a comparação de três quadros consecutivos, definindo uma medida de dissimilaridade denominada *Observer Motion Coherence (OMC)*.

Wu et. al. [WWL98] definem a projeção do desvio padrão da diferença entre pixels para detectar ocorrências de alguns tipos de wipes em vídeos, no domínio de compressão do MPEG. Essa projeção foi definida como uma medida estatística utilizada para detectar platôs monodimensionais que determinam a ocorrência de wipes no vídeo para a medida escolhida. Em caso de ocorrências de rápidos movimentos de objetos, os autores sugerem a utilização de uma checagem adicional para cada quadro anteriormente detectado, definidos agora como quadro candidato a ocorrência do wipe.

Uma variação da comparação dos pixels entre quadros consecutivos, também apresentada em [KC01, ZKS93], utiliza a quantidade de pixels alterados entre os quadros f_t e f_{t-1}, comparada com um determinado limiar T_1. De acordo com as Equações 2.3 e 2.4, um corte será caracterizado se o percentual D_t for maior que um limiar T_2.

$$C_t(x, y) = \begin{cases} 1, & \text{se } |f_{t-1}(x, y) - f_t(x, y)| > T_1 \\ 0, & \text{caso contrário} \end{cases} \tag{2.3}$$

$$D_t = \frac{1}{W * H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} C_t(x, y) \tag{2.4}$$

Mesmo com essa modificação, essa abordagem ainda é sensível a trabalhos de câmera como *panning* ou a movimentos de objetos nos quadros. [KC01] apresenta como alternativa para reduzir parcialmente este problema a aplicação de um filtro médio (*smoothing filter*) [GW92] antes das comparações entre pixels de dois quadros consecutivos.

Comparação de Histogramas

Algumas abordagens utilizam o histograma dos quadros objetivando agregar mais informações de cor do que informações espaciais. Um dos primeiros trabalhos baseados em histograma substitui a diferença entre os pixels dos quadros consecutivos na Equação 2.1, pela diferença dos histogramas desses quadros, conforme a Equação 2.5 [KC01]. A utilização da diferença entre histogramas não é suficiente para os casos em que dois quadros cujos conteúdos sejam completamente diferentes possuam o mesmo histograma. Para esses casos, Pass e Zabih [PZ99] apresentaram uma adaptação do histograma, denominado *joint histogram*, que considera este problema.

$$D_t = \frac{1}{L} \sum_{j=0}^{L-1} |H_{ist_{t-1}}(j) - H_{ist_t}(j)| \tag{2.5}$$

Outras abordagens aplicam medidas estatísticas sobre o histograma [BMM99, KC01, NPC98, PS97, SP95]. Sethi e Patel [SP95], por exemplo, verificaram a utilização de três testes estatísticos para a comparação entre histogramas, no domínio de compressão MPEG: o teste do χ^2, o teste da taxa de propabilidade de Yakimovsky [Yak76], e o teste do Kolmogorov-Smirnov [GSF77]. Segundo os autores, os melhores resultados foram alcançados utilizando-se o teste do χ^2. Apesar do aumento do tempo computacional, a utilização do teste do χ^2 permite realçar a diferença entre os histogramas dos quadros consecutivos. Este teste, aplicado aos histogramas H_{ist_t} e $H_{ist_{t-1}}$ para os quadros f_t e f_{t-1}, respectivamente, é dado pela Equação 2.6. Porém, além de realçar a diferença na presença de cortes, o teste do χ^2 realça ainda a diferença entre quadros com presença de movimentos de objetos e câmera, resultando em falsas detecções.

$$D_t = \sum_{j=0}^{L-1} \frac{(H_{ist_{t-1}}(j) - H_{ist_t}(j))^2}{H_{ist_t}(j)} \tag{2.6}$$

Outras abordagens considerando a intersecção de histogramas e a diferença ponderada em relação a uma determinada vizinhança de pixels podem ser encontrados em [KC01, NPC98]. Uma outra medida semelhante à intersecção da diferença entre histogramas é apresentada em [ZWB+99].

2.1.2 Informações de Textura e Forma

Segundo [GW92], não existe nenhuma definição formal para textura. Podemos dizer que a textura é caracterizada pela distribuição espacial da intensidade dos pixels em uma determinada vizinhança [AG03]. Alguns trabalhos consideram a transformada de *wavelet* para a segmentação de texturas em imagens [AG03]. Por considerar informações espaciais, alguns métodos utilizam essa abordagem para definir uma medida de dissimilaridade entre os quadros de um segmento de vídeo digital. Mandal e Panchanathan [MP01] exploram o uso de wavelet no domínio da compressão, analisando vários trabalhos relacionados e definindo algoritmos para a segmentação de vídeo, incluindo a detecção de cortes e de transições graduais.

Além da textura, características de formas podem ser utilizadas para a segmentação estrutural de um vídeo. Os objetos presentes nos diversos quadros são utilizados, na maioria das vezes, para a obtenção de arestas originadas pelos objetos. As arestas são definidas nas regiões onde ocorrem mudanças drásticas em intensidades de cor. Heng e Ngan [HN01b] apresentam uma técnica para detecção de transições abruptas e graduais, baseada em informações de objetos. Eles definem um objeto aresta que representa uma coleção de pixels com algumas informações associadas, tais como, orientação, direção e tempo de vida durante uma tomada. O algoritmo analisa o tempo de vida de cada objeto aresta para definir as ocorrências das transições, mesmo que estas sejam longas, como no caso das transições graduais. Apesar dos autores apresentarem bons resultados para o método proposto, o algoritmo tem custo computacional alto e apresenta limitações na ocorrência de certos efeitos de câmera, tais como flash e panning. Song e Ra [HN01a] apresentam também uma abordagem para detecção de tomadas baseada em arestas. O algoritmo proposto, aplicado sobre vídeo no domínio da compressão MPEG, explora as características das imagens DC, para a obtenção das arestas, utilizando um tempo computacional menor e reduzindo a quantidade de memória necessária para este processo. Um segundo passo do algoritmo utiliza um método de clusterização que considera, além das arestas encontradas nos quadros, a diferença dos pixels e dos histogramas entre quadros consecutivos, na obtenção dos quadros onde ocorrem transições, tanto cortes quanto fades e dissolves.

2.1.3 Comparação por Blocos

Alguns trabalhos utilizam técnicas semelhantes às discutidas anteriormente porém, algumas modificações são realizadas com o objetivo de agregar informações espaciais às medidas utilizadas. Cada quadro é dividido em blocos (subconjuntos do quadro) de tamanhos variados, que são utilizados nas comparações entre quadros. Em [NT91], Nagasaka e Tanaka analisam várias técnicas envolvendo comparação por blocos, utilizando medidas

de comparação entre pixels e histogramas. Segundo os autores, os melhores resultados acontecem quando a partição de cada quadro é realizada de forma que se tenha 16 blocos de mesmo tamanho, e o teste do χ^2 aplicado nos histogramas dessas regiões, descartando as grandes diferenças, permitindo reduzir o efeito da presença de movimentos de objetos e câmera.

Dugad et. al. [DRA98] utilizam apenas a divisão por blocos na segunda fase de seu algoritmo. Primeiramente, a diferença entre histogramas é utilizada e comparada com um limiar. Para os quadros cujos valores são inferiores ou iguais ao limiar, divide-se os quadros em blocos de tamanhos pré-definidos e realiza-se a comparação utilizando a medida estatística da taxa de probabilidade. Este valor é comparado com um segundo limiar para completar a detecção de tomadas.

Em [DB99], Demarty e Beucher apresentam uma abordagem em que quadros do vídeo são divididos em pequenos blocos, em geral de 20x20 pixels, e a distância média euclidiana entre os blocos de quadros consecutivos é calculada para a criação da chamada máscara de transição geométrica. Um vetor monodimensional é criado com a soma de todos os valores de cada máscara e filtros morfológicos são aplicados para a detecção das transições, utilizando um limiar predefinido. Outras técnicas utilizando comparação por blocos podem ser encontradas em [BMM99, LYL01, Gui03].

2.1.4 Twin Comparison

Este método introduzido em [BMM99, KC01, Gui03], considera que quadros de diferentes tomadas têm o valor de dissimilaridade maior do que quadros da mesma tomada, para alguma medida de dissimilaridade escolhida, e que transições graduais, em que os primeiros quadros não são totalmente diferentes da tomada anterior, podem ser detectados se dois limiares forem utilizados. Basicamente, um limiar L_s é definido para a detecção de cortes e um segundo limiar $L_d < L_s$ é considerado na detecção das transições graduais. De acordo com [ZKS93], se a medida de dissimilaridade D_t entre os quadros no instante t e $t-1$ satisfizer a condição $L_d < D_t < L_s$, então o quadro f_t é definido como candidato a um quadro inicial de transição gradual. Para cada quadro candidato f_t, o valor acumulado $A(j) = \sum D_j$ é calculado para todo quadro no instante $j \geq t$ e $D_j < L_s$. Quando $A(j) > T_s$, uma transição gradual é assumida entre os quadros t e j.

2.1.5 Clusterização Temporal

A maioria das técnicas que utiliza alguma medida de dissimilaridade define um ou mais limiares para a detecção das transições. Entretanto, essa definição pode variar conforme o tipo do vídeo a ser analizado. Para resolver este problema, pode-se aplicar algum algoritmo de clusterização automático, que irá definir o limiar segundo as características

da medida de dissimilaridade escolhida, independente do tipo do vídeo. Isso é possível porque a segmentação temporal do vídeo pode ser vista como um problema de duas classes de clusters (corte e não corte), como descrito em [KC01].

Em [FT98], Ferman e Tekalp definem um vetor de diferenças entre pixels e outro vetor de diferenças de histogramas, formando um espaço bidimensional. Após o cálculo destas diferenças, um filtro mediano [GW92] é aplicado para a eliminação de ruídos em ambos os vetores. Utilizando o espaço bidimensional criado com os eixos da diferença entre pixels e da diferença entre histogramas, um método de clusterização é aplicado para a obtenção dos quadros em que transições abruptas acontecem. Uma variação desse método é proposta para a extração de quadros-chaves, e uma outra é utilizada na detecção de cortes durante o processamento do vídeo.

Outro método que considera a diferença entre pixels e a diferença entre histogramas para a detecção de cortes é apresentada por Krinidis et. al. em [KTP01]. Ao mesmo tempo, é utilizada a análise de áudio do segmento de vídeo, classificando este sinal em fala, silêncio e música, com o objetivo de permitir a detecção de cenas. Esse conjunto promoveu um médio resultado para a detecção de cenas, mas introduziu um importante fator, a análise de áudio, que poderá contribuir ainda mais para a segmentação estrutural de vídeo digital.

2.1.6 Informações de Movimentos

As informações dos vetores de movimentos (*motion vectors*) dos pixels representam uma das vantagens de se analisar o vídeo digital no domínio da compressão MPEG. Esses vetores representam a direção predominante do movimento de um determinado grupo de pixels, utilizado para compensação de movimento entre quadros comprimidos no padrão MPEG. Durante uma transição abrupta, espera-se que os vetores de movimentos não tenham nenhuma direção predominante. Alguns trabalhos baseiam-se neste aspecto para detectar cortes. Esta informação pode ser considerada ainda na detecção de efeitos de câmera, tais como zoom-out, zoom-in e panning.

Por exemplo, Patel97 e Sethi [PS97] utilizam esses vetores de movimentos para a identificação de efeitos de câmera nas tomadas, realizando a classificação das mesmas em seis categorias conforme o tipo do efeito. Outros trabalhos baseados em vetores de movimentos, tanto para detecção de transições como para efeitos de câmera, podem ser encontrados em [BMM99, KC01, Gui03].

2.2 Imagem Representativa

Algumas abordagens utilizam uma imagem representativa do vídeo para a detecção de alguns de seus eventos. Essa imagem representativa varia conforme a técnica utilizada e um exemplo significativo da mesma refere-se à representação denominada **ritmo visual** apresentada mais adiante.

Akutsu e Tonomura [AT94] definem o Método de Vídeo por Tomografia que utiliza duas imagens, raio-x (*x-ray*) e raio-y (*y-ray*), como uma transformação do vídeo digital capaz de fornecer padrões de identificação de efeitos de câmera. Essa transformação é realizada fixando-se a variável y como constante, para a obtenção da imagem raio-x. Analogamente, fixando-se x como constante obtém-se a imagem raio-y. Em seguida, os autores aplicam um filtro de arestas nessas imagens e utilizam a transformada de Hough para a obtenção de alguns efeitos de câmera. Joly e Kim apresentaram em [JK96] uma simplificação desse método, eliminando o filtro de arestas e preservando a transformada de Hough, melhorando o desempenho do algoritmo.

Em [LYL01], Lee et. al. aplicam a transformada rápida de Fourier sobre as imagems raio-x e raio-y e, em seguida, utilizam uma rede neural para a detecção de efeitos de câmera, tais como, movimentos horizontais e verticais da câmera, zoom-out e zoom-in.

Ritmo Visual

Mais recentemente, novas técnicas de segmentação estrutural de vídeo foram definidas com base em uma imagem representativa do segmento de vídeo. Esse modelo de representação é denominado **Espaço-Temporal** (*Spatial-Temporal*) [NPC99a, NPC99b, NPC00, NPZ03], ou ainda **Ritmo Visual** (*Visual Rhythm*) [CLK$^+$99, KLY$^+$01, GCLdAA01, Gui03]. Essa dissertação refere-se a esta representação apenas como ritmo visual. A maioria dos trabalhos que utiliza este modelo procura reduzir a quantidade de dados a serem analisados, reduzindo o tempo computacional para a segmentação do vídeo, além de utilizar técnicas já conhecidas de processamento de imagens sobre a representação. Algumas propostas de sub-amostragem dos quadros de um vídeo foram feitas com o objetivo de aumentar a eficiência dos algoritmos [ZKS93]. No entanto, a obtenção de bons resultados está diretamente associada à escolha do tamanho das sub-amostragens espacial e temporal que podem variar conforme o vídeo utilizado. O ritmo visual realiza uma sub-amostragem dos quadros preservando grande parte das características do segmento de vídeo digital.

Esta representação pode ser entendida como uma projeção do vídeo do domínio $2\mathbb{D}$ + t para $\mathbb{D}$ + t, transformando a segmentação estrutural de vídeo, a partir dos trabalhos de [GCLdAA01, GLCdAA02, GLCdAA03b, GLCdAA03a, Gui03, BL03], em um problema de segmentação de imagens. Essa projeção, em linhas gerais, pode ser definida da seguinte forma:

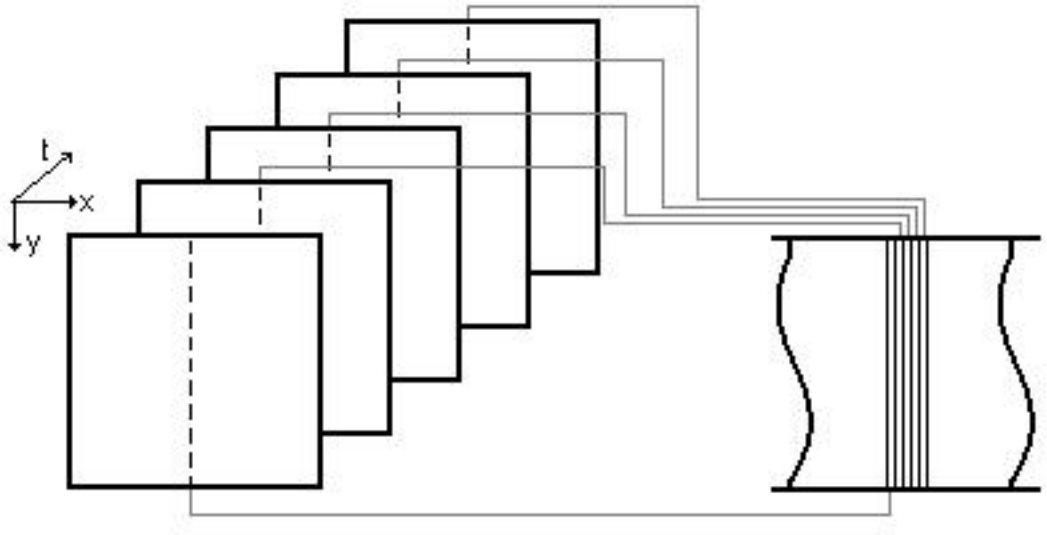


Figura 2.2: Definição do ritmo visual utilizando uma linha vertical central.

Definição 2.2. Ritmo Visual - *Seja V um segmento de vídeo digital de N quadros (Definição 1.2), no domínio $2\mathbb{D} + t$. Sejam H e W, respectivamente, a altura e a largura de cada frame f_t desse vídeo. O ritmo visual é a transformação de cada frame f_t em uma coluna da imagem de ritmo visual VR definido como:*

$$VR(t, z) = f_t(r_x \times z + a, r_y \times z + b)$$

em que $z \in [0, H_{VR} - 1]$ e $t \in [0, N - 1]$, H_{VR} e N são, respectivamente, a altura e a largura da imagem de ritmo visual; r_x e r_y representam a taxa de amostragem dos pixels e a e b, o deslocamento em cada frame.

Note que a largura da imagem de ritmo visual é exatamente igual à quantidade de frames do segmento de vídeo digital. Para exemplificar, de acordo com os parâmetros nesta definição, pode-se obter uma sub-amostragem do vídeo considerando apenas a diagonal principal, se $a = b = 0$, $r_x = r_y = 1$; ou uma linha vertical central (Figura 2.2), se $r_x = b = 0$, $r_y = 1$ e $a = H/2$; ou ainda uma linha horizontal central, se $r_y = a = 0$, $r_x = 1$ e $b = W/2$. A Figura 2.3 ilustra um exemplo de ritmo visual de um segmento de vídeo digital em que $r_x = b = 0$, $r_y = 1$ e $a = H/2$.

Observa-se que essa projeção do segmento de vídeo para a imagem representativa do ritmo visual, composta pela correlação entre espaço e tempo do segmento de vídeo, indica a localização dos quadros onde as transições acontecem. Em geral, a imagem de ritmo visual torna visualmente perceptível o corte como uma linha separadora vertical

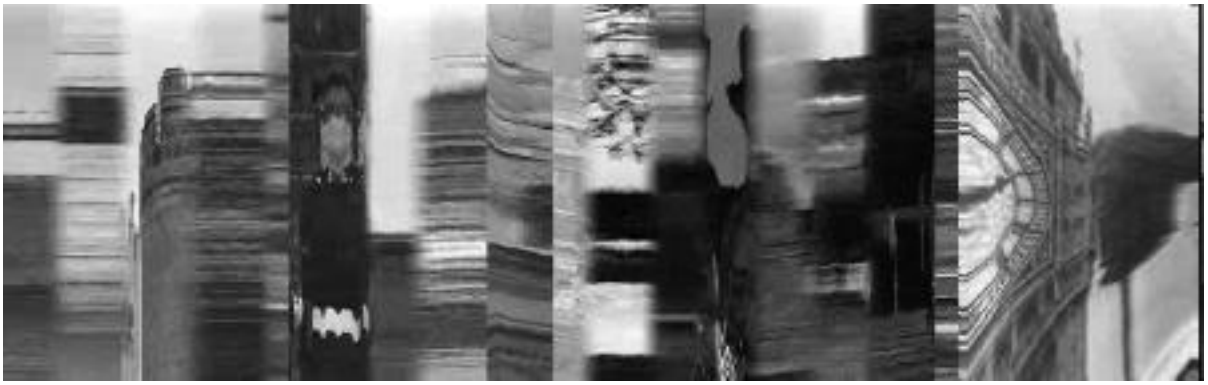


Figura 2.3: Exemplo do ritmo visual de um segmento de vídeo digital.

na imagem; um wipe como uma linha divisória inclinada (dependendo do tipo de wipe presente e da amostragem escolhida para o ritmo visual) e um dissolve como duas regiões conectadas por uma "nuvem" sem definição de separação [NPC99b]. Neste caso, a classificação do tipo de transição utilizando ritmo visual requer, na maioria das vezes, as três principais componentes de cada quadro, resultando na utilização das imagens de ritmo visual referente às linhas da vertical, da horizontal e de uma das diagonais.

C. W. Ngo et. al. [NPC99a] propõem a utilização desse modelo no domínio da compressão MPEG, para a detecção de cortes e alguns padrões de wipes, considerando a imagem projetada. O modelo proposto permite classificar cortes e wipes, além de identificar a duração das seqüências de wipes, e utiliza duas imagens de ritmo visual: uma composta pela linha central horizontal de cada quadro e outra, pela linha central vertical. A combinação dessas duas imagens fornece a classificação do tipo de transição e a identificação e duração do tipo de wipe detectado. O algoritmo utiliza segmentação da imagem por cadeias de Markov, procurando descontinuidades de textura e cor nas regiões de transições.

Posteriormente, C. W. Ngo et. al. [NPC99b] aprimoraram o método anterior para permitir a detecção de dissolves. Passaram então a utilizar três imagens de ritmo visual: uma para a linha vertical central de cada quadro, outra para a linha horizontal central e uma terceira composta pela diagonal principal. A utilização das três imagens de ritmo visual possibilita a detecção de alguns tipos de wipes. Com um algoritmo ainda mais robusto, C. W. Ngo et. al. [NPC00] readaptaram seu modelo de detecção de wipes para permitir a identificação de outros padrões, com redução da sensibilidade a efeitos de câmeras, ainda utilizando características de textura e cor da imagem do ritmo visual.

Em trabalho recente, Ngo et. al. [NPZ03] utilizam a representação do ritmo visual para a análise e segmentação de vídeo (separação de objetos e fundo) baseadas na análise

de movimentos, identificando alguns efeitos de câmera em vídeos de esportes. O método proposto baseia-se na formulação de estruturas de tensores, e o ritmo visual utilizado para a análise dos padrões de efeitos de câmera é obtido no domínio da compressão MPEG.

M. G. Chung et. al. [CLK+99] consideram sobre as imagens de ritmo visual, originadas a partir do vídeo comprimido em MPEG, a média local e a variância do gradiente horizontal na detecção de cortes. Após a remoção dos cortes, o algoritmo utiliza a medida de número de pixels conectados, realçando as regiões limitantes das ocorrências de wipes. Em seqüência, o algoritmo considera a taxa de variação da luminância ao longo dos quadros na detecção de dissolves e fades.

Drew et. al. [DLZ00] utilizaram uma variação do ritmo visual, ainda sobre o domínio da compressão MPEG, que considera a utilização do histograma cromático de cada quadro na criação da imagem de ritmo visual, atribuindo uma informação global, diferentemente das outras propostas. Utilizando o espaço Cb e Cr do espaço de cores YCbCr, e a medida de diferença de histogramas, o método proposto identifica wipes e dissolves nos segmentos de vídeo.

Guimarães et. al. [GCLdAA01] desenvolveram um algoritmo para detecção de cortes, fora do domínio da compressão MPEG, realizando uma etapa de filtragem através de operadores morfológicos [Ser82] como abertura e fechamento. O método considera ainda o gradiente horizontal na imagem de ritmo visual, e uma operação de afinamento aplicada aos picos definidos pelo gradiente horizontal e, posteriormente, a detecção de seus pontos maximais. Os autores propuseram uma variação deste algoritmo para a detecção de flashes. Posteriormente, apresentaram em [GLCdAA02] um algoritmo multiescala parametrizado baseado em resíduos morfológicos [LG01], para a detecção de transições abruptas (corte) e graduais (dissolve), utilizando vários operadores morfológicos e granulometria. Uma outra variação utilizada pelos mesmos autores em [GLCdAA03b, GLCdAA03a, Gui03] considera, para a criação do ritmo visual, o histograma em níveis de cinza de cada quadro, definindo cada histograma como uma coluna da imagem de ritmo visual. A partir dessa informação mais global, os autores definem algoritmos para detecção de cortes, fades, dissolves e flashes, utilizando filtros morfológicos, gradiente e afinamento, de forma semelhante à apresentada em [GCLdAA01].

Kim et. al. propuseram em [KLY+01] a utilização do ritmo visual para a verificação de tomadas. Os autores definem um ambiente para auxiliar um operador a identificar e validar as transições em um vídeo digital, utilizando a imagem de ritmo visual como guia de consulta das transições, além de permitir a identificação e classificação de efeitos de câmera.

Recentemente, Bezerra e Leite [BL03] apresentaram uma nova abordagem aplicando casamento de strings entre as colunas da imagem de ritmo visual. O método utiliza o algoritmo LCS (*Longest Commum Substring*) [CLRS02], e os valores obtidos são armaze-

nados em um vetor utilizado na detecção das transições. Desta forma, a segmentação do vídeo passa a ver vista como um problema de segmentação do sinal $1\mathbb{D}$. Alguns resultados preliminares foram apresentados para a detecção de cortes e dissolves.

O próximo capítulo descreve o método utilizado nesta dissertação para a detecção de cortes e flashes.

Capítulo 3

Detecção de Cortes e Flashes: Uma Nova Abordagem

Como discutido no Capítulo 2, várias abordagens foram propostas para a detecção de transições abruptas (cortes) utilizando as medidas de dissimilaridade. Muitas delas requerem um alto custo computacional e são, em geral, sensíveis à presença de movimentos de objetos e câmera. Outras abordagens baseadas em uma imagem representativa do vídeo, aplicam métodos de processamento de imagens que, em sua maioria, requerem vários parâmetros e algoritmos sofisticados, de elevado custo de processamento. Este capítulo descreve os detalhes de uma metodologia empregada na detecção de transições e eventos abruptos, envolvendo cortes e flashes, a partir de algoritmos simples e eficientes.

3.1 Detecção de Cortes

A utilização da representação do ritmo visual por amostragem de um segmento de vídeo reduz significativamente a quantidade de dados analisados numa seqüência, porém, essa representação considera apenas informações regionais de cada quadro, definidas de acordo com os parâmetros utilizados na criação dessa nova imagem. Para se considerar informações globais dos quadros do segmento de vídeo, algumas abordagens ao modelo do ritmo visual foram propostas.

Inicialmente, o histograma em níveis de cinza (Definição 2.1) de cada quadro foi utilizado com o objetivo de agregar informações globais ao ritmo visual. A característica de invariância à rotação e translação do histograma representa uma das vantagens do seu emprego como representação global dos quadros do vídeo. No entanto, é sabido que dois quadros totalmente diferentes podem ter o mesmo histograma, fato este que dificilmente ocorre nas transições dos segmentos de vídeo. De certa forma, este problema pode ser atenuado com a utilização do *joint histogram* [PZ99] ao invés do histograma convencional.

O **ritmo visual por histograma** pode ser definido da seguinte forma:

Definição 3.1. Ritmo Visual por Histograma - *Seja H_{ist_t} o histograma do t-ésimo quadro de um vídeo digital V qualquer, com N quadros. O ritmo visual por histograma $VR_{hist}(t, z)$ é dado por:*

$$VR_{hist}(t, z) = H_{ist_t}(z),$$

para $z \in [0, H_{VR} - 1]$ e $t \in [0, N - 1]$. H_{VR} e N são, respectivamente, a altura e a largura da imagem de ritmo visual por histograma. De modo geral, H_{VR} representa o número de níveis de cinza ou intensidades de cores utilizadas.

Guimarães et. al. apresentam em [GLCdAA03b, GLCdAA03a, Gui03] métodos que utilizam o ritmo visual por histograma na detecção de transições abruptas e graduais em segmentos de vídeo digital. A Figura 3.1 ilustra o ritmo visual por histograma para o segmento apresentado na Figura 2.3. Apenas para efeito de visualização, esta figura foi normalizada para 256 níveis de cinza.

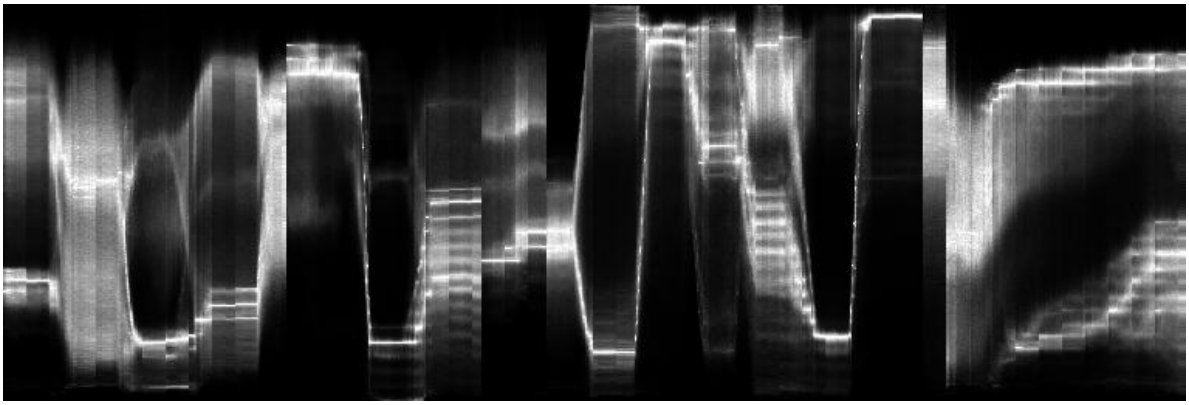


Figura 3.1: Exemplo de ritmo visual por histograma de um segmento de vídeo digital.

Nas próximas seções, serão apresentadas e discutidas as etapas do método aqui proposto para detecção de transições abruptas.

3.1.1 Imagem Representativa

Este trabalho considera o histograma acumulado de uma imagem (Definição 3.2, a seguir) como uma alternativa para agregar informações globais dos quadros que compõem uma seqüência de vídeo. Este histograma é invariante à rotação e translação e pode ser empregado na definição de uma imagem de ritmo visual, de acordo com a Definição 3.3, apresentada posteriormente.

Definição 3.2. Histograma Acumulado - *Seja H_{ist} o vetor de histograma de uma imagem ou de um quadro, cuja intensidade de cores varia no intervalo $[0, L-1]$. Seja j uma intensidade qualquer de cor pertencente a este mesmo intervalo. O histograma acumulado $H_{acum}(j)$ é dado por:*

$$H_{acum}(j) = \sum_{i=0}^{j} H_{ist}(i)$$

Definição 3.3. Ritmo Visual por Histograma Acumulado - *Seja H_{acum_t} o histograma acumulado do t-ésimo quadro de um vídeo digital V qualquer, com N quadros. O ritmo visual por histograma acumulado $VR_{ha}(t, z)$ é dado por:*

$$VR_{ha}(t, z) = H_{acum_t}(z),$$

para $z \in [0, H_{VR} - 1]$ e $t \in [0, N-1]$. H_{VR} e N são, respectivamente, a altura e a largura da imagem de ritmo visual por histograma acumulado. De modo geral, H_{VR} também representa o número de níveis de cinza ou intensidades de cores utilizadas.

A Figura 3.2 exemplifica o VR_{ha} para o mesmo segmento de vídeo da Figura 3.1. Nesta figura, percebe-se a existência de certas linhas de contraste verticais que caracterizam a ocorrência de transições abruptas no vídeo digital. Diferentemente do ritmo visual por histograma, o ritmo visual por histograma acumulado permite distinguir melhor este contraste, principalmente nas regiões dos quadros entre as transições. Para efeito de exibição, a freqüência dos intervalos de cores do histograma acumulado, presentes na imagem do VR_{ha} é normalizada para 256 níveis de cinza.

O VR_{ha} é considerado, neste trabalho, como a imagem representativa a ser explorada na detecção de transições abruptas de uma seqüência de vídeo. O baixo custo computacional para o cálculo do histograma acumulado e construção da imagem representativa correspondente é um fator importante no desempenho do método, que considera apenas uma leitura do vídeo na construção dessa imagem cujo processamento passa a não depender de novos acessos ao vídeo original.

3.1.2 Vetor Representativo

De um modo geral, a imagem representativa no formato de um ritmo visual oferece uma boa redução na quantidade de dados a serem analisados. Visando uma melhor caracterização da informação das transições abruptas contida numa imagem do VR_{ha}, considera-se aqui uma simplificação ainda maior da informação presente nesta imagem, através de uma projeção vertical da mesma que define um vetor representativo do vídeo digital da forma:

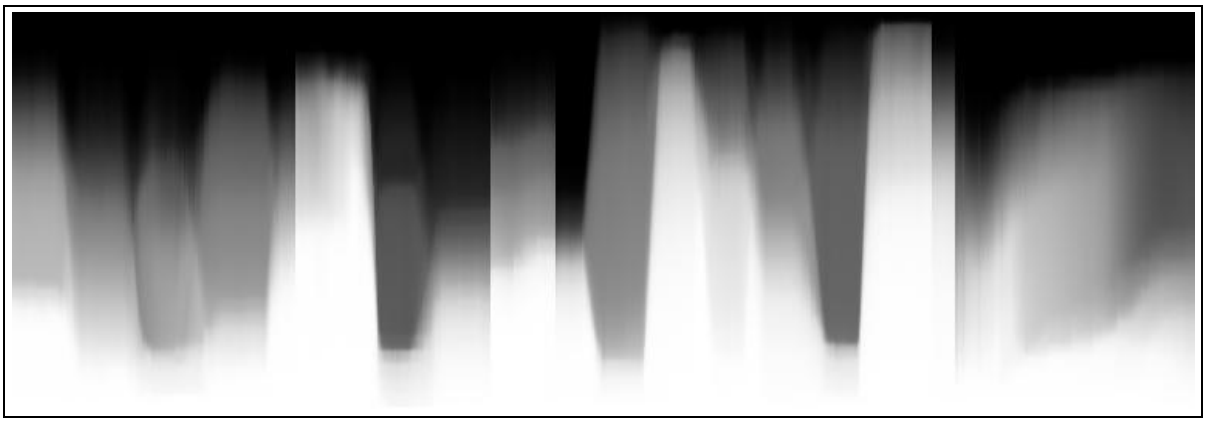


Figura 3.2: Exemplo de VR_{ha} de um segmento de vídeo digital.

Definição 3.4. Projeção Vertical - *Seja Im uma imagem de altura H e largura W. O vetor $\mathcal{P}$ da projeção de cada coluna t da imagem Im pode ser definido como:*

$$\mathcal{P}(t) = \sum_{j=0}^{H-1} Im(j,t),$$

para $t \in [0, W-1]$, em que t representa a t-ésima coluna da imagem Im e j a j-ésima linha.

A Figura 3.3 ilustra a projeção do VR_{ha} da Figura 3.2. Uma análise do "relevo" do vetor de projeção vertical, apresentado na Figura 3.3, permite identificar certos "degraus" referentes aos quadros associados a transições abruptas na seqüência. Observe que o problema de detecção de transições abruptas é transformado aqui num problema de baixo custo computacional, constituindo-se na detecção de padrões num sinal $1\mathbb{D}$.

3.1.3 Laplaciano

A partir das considerações anteriores sobre os degraus significativos no relevo do vetor representativo $\mathcal{P}$, indicando a presença de cortes num vídeo digital, o passo seguinte na identificação destes degraus consiste da aplicação de um laplaciano sobre $\mathcal{P}$ para realçar a ocorrência dessas transições e minimizar a contribuição das variações graduais no mesmo vetor. Este laplaciano pode ser definido da seguinte forma:

Definição 3.5. Laplaciano - *Seja $\mathcal{P}$ um vetor de tamanho W. O vetor laplaciano $\mathcal{L}_{ap}$, definido sobre $\mathcal{P}$, pode ser representado por:*

$$\mathcal{L}_{ap}(t) = |2 * \mathcal{P}(t) - \mathcal{P}(t-1) - \mathcal{P}(t+1)|,$$

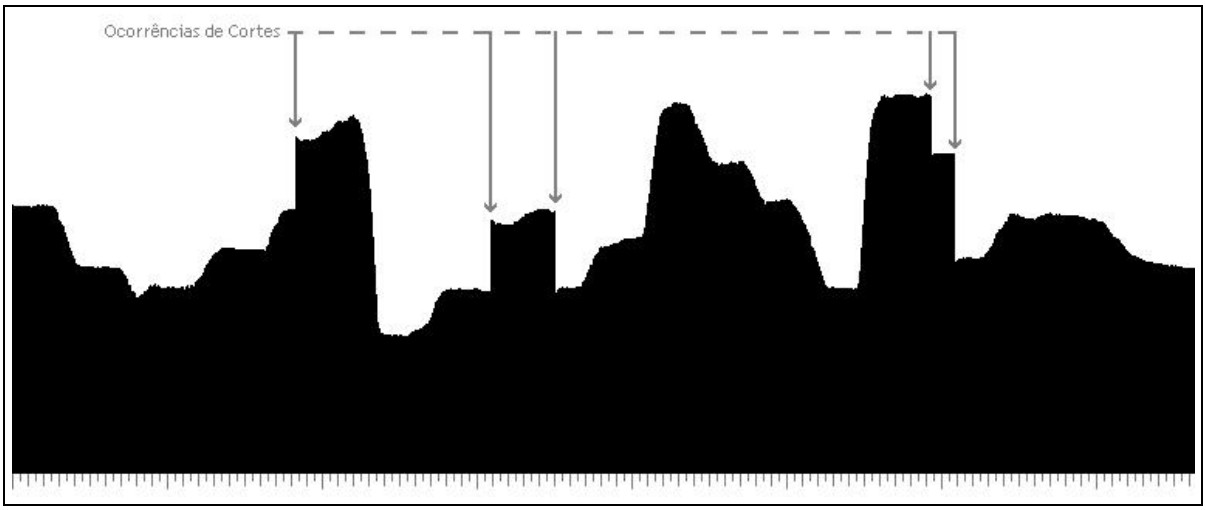


Figura 3.3: Exemplo de projeção vertical do VR_{ha} de um segmento de vídeo digital.

para $t \in [1, W-2]$.

Assim, variações abruptas em $\mathcal{P}$ resultam em valores elevados em $\mathcal{L}_{ap}$, enquanto que variações graduais, semelhantes a uma rampa resultam em baixos valores neste vetor. Observe que a definição do laplaciano aqui considerada não faz distinção entre valores positivos e negativos na resposta (para se evitar um deslocamento na posição real dos cortes na seqüência).

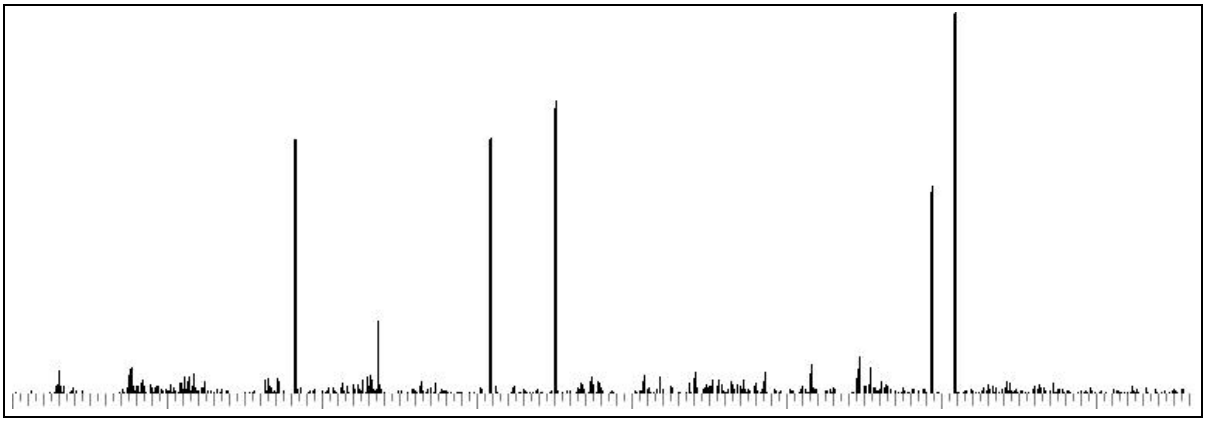


Figura 3.4: Laplaciano da projeção vertical do VR_{ha} apresentado na Figura 3.3.

3.1.4 White Top-hat

Naturalmente, o vetor $\mathcal{L}_{ap}$ contém, neste ponto, vários picos no sinal $1\mathbb{D}$ associados aos quadros de transições abruptas no vídeo digital. Um realce adicional destes picos pode ser considerado através de um simples operador morfológico denominado chapéu mexicano (*white top-hat*) [Ser82], definido a seguir.

Definição 3.6. White Top-hat - *O operador white top-hat WTH de tamanho n representa o resíduo de uma abertura morfológica de mesmo tamanho, da forma:*

$$WTH_n = Id - \gamma_n,$$

em que Id é o operador identidade e γ_n é uma abertura morfológica de tamanho n dada pela aplicação sucessiva de n erosões e dilatações sobre o sinal com uma função estruturante elementar $\mathbf{b}$

Informalmente, o WTH_n preserva as componentes de um sinal de dimensão menor que a função estruturante especificada.

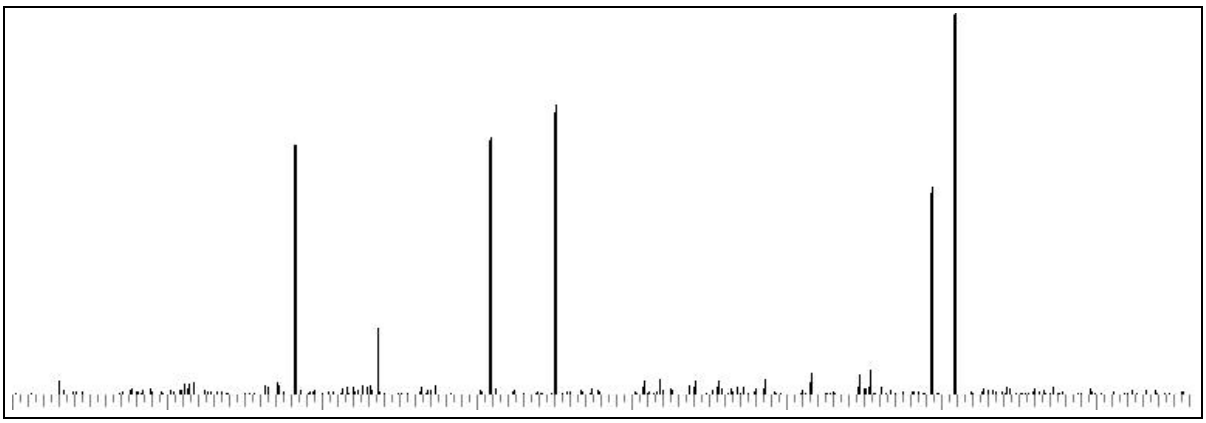
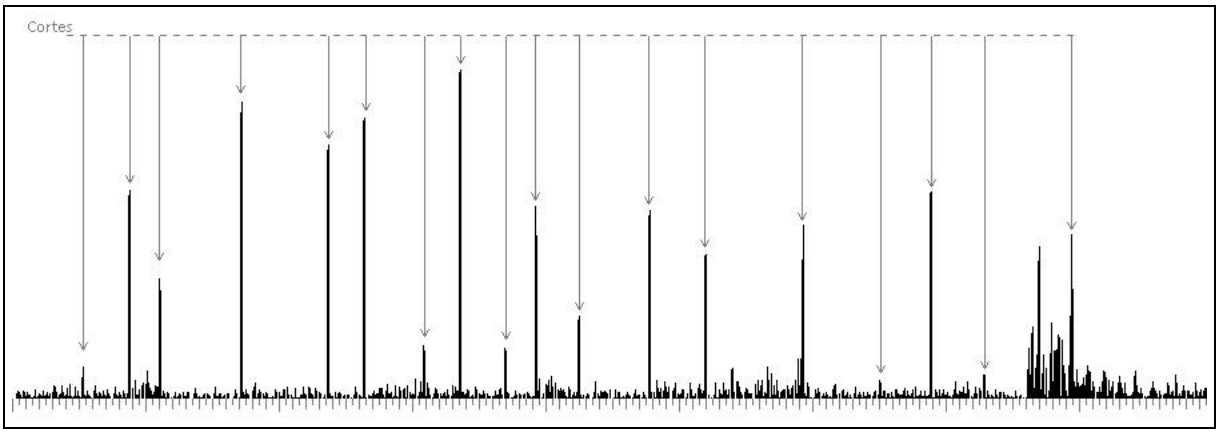
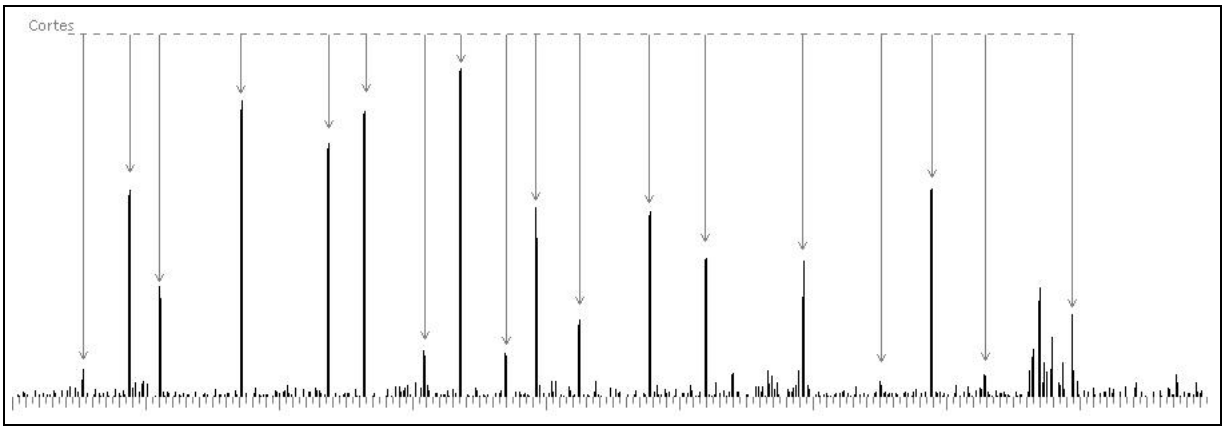


Figura 3.5: White top-hat aplicado sobre o sinal da Figura 3.4.

A função estruturante b depende da quantidade de quadros envolvida na transição. Como as transições abruptas acontecem entre dois quadros consecutivos, uma função estruturante discreta, planar e simétrica de dimensão 1×3 é suficiente para manter os picos significativos do sinal e garantir uma filtragem do mesmo. A Figura 3.5 ilustra a aplicação do WTH_3 no vetor laplaciano apresentado na Figura 3.4. Um outro sinal justificando melhor a aplicação do WTH_3 é apresentado na Figura 3.6. Esta figura refere-se ao laplaciano da projeção vertical de um vídeo de comercial contendo 18 cortes.



a) Laplaciano da projeção vertical do VR_{ha}.



b) White top-hat aplicado sobre a).

Figura 3.6: Laplaciano e white top-hat sobre o VR_{ha} de um vídeo de comercial.

3.1.5 Limiarização Versus Classificação

Nesta etapa do processamento, uma forma simples e manual de se detectar os cortes num segmento de vídeo é através de uma limiarização do vetor representativo $\mathcal{P}$, após a aplicação do laplaciano e do WTH_3. Neste caso, a qualidade da resposta depende naturalmente do valor do limiar escolhido: um valor baixo pode gerar falsas detecções enquanto que um valor alto, perdas nesta detecção. Além disso, esta escolha pode variar de acordo com as características de cada vídeo.

Uma forma de se automatizar este processo e torná-lo independente do tipo de vídeo analisado é através de métodos de classificação/clusterização em que duas classes (uma representando transições abruptas e a outra, os demais tipos de eventos) são identificadas. Este trabalho considera o método de clusterização K-means [KMN+00], como será visto a seguir.

K-means

O K-means é um método de clusterização iterativo, em que K centros, definindo K-classes são considerados, para um conjunto qualquer de dados. O algoritmo implementado nesta dissertação, apresentado a seguir, utiliza dois centros, sendo executado no máximo $MaxI$ iterações. Observou-se que o número máximo de iterações realizadas pelo algoritmo não foi superior a 10 iterações. Dessa forma, definiu-se $MaxI = 30$ como margem de segurança.

Algoritmo 3.1 K-means (K-means para dois centros)

Entrada: vetor de dados V
Saída: dois grupos: $G1$ e $G2$

1: $centro1 \leftarrow$ Menor elemento de V;
2: $centro2 \leftarrow$ Maior elemento de V;
3: $execucao \leftarrow 0$;
4: **repeat**
5: **for all** elemento $e \in V$ **do**
6: **if** $|e - centro1| \geq |e - centro2|$ **then**
7: $G2 \leftarrow G2 \cup \{e\}$;
8: **else**
9: $G1 \leftarrow G1 \cup \{e\}$;
10: $centro1 \leftarrow$ Média dos elementos de $G1$;
11: $centro2 \leftarrow$ Média dos elementos de $G2$;
12: $execucao \leftarrow execucao + 1$;
13: **until** (($centro1$ não foi alterado) and ($centro2$ não foi alterado)) or ($execucao \geq MaxI$)
14: return $\langle G1, G2 \rangle$;

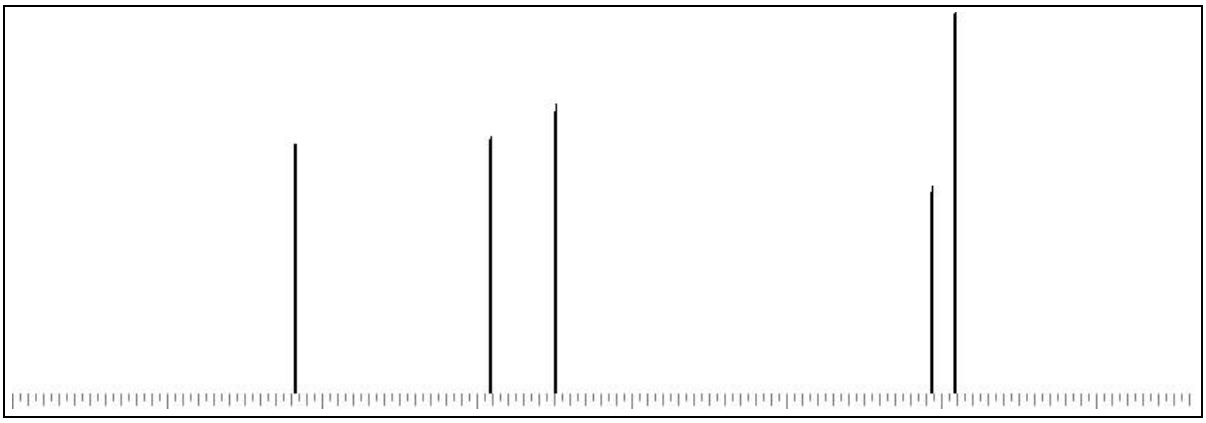


Figura 3.7: Aplicação do K-means (2 classes) sobre o sinal da Figura 3.5.

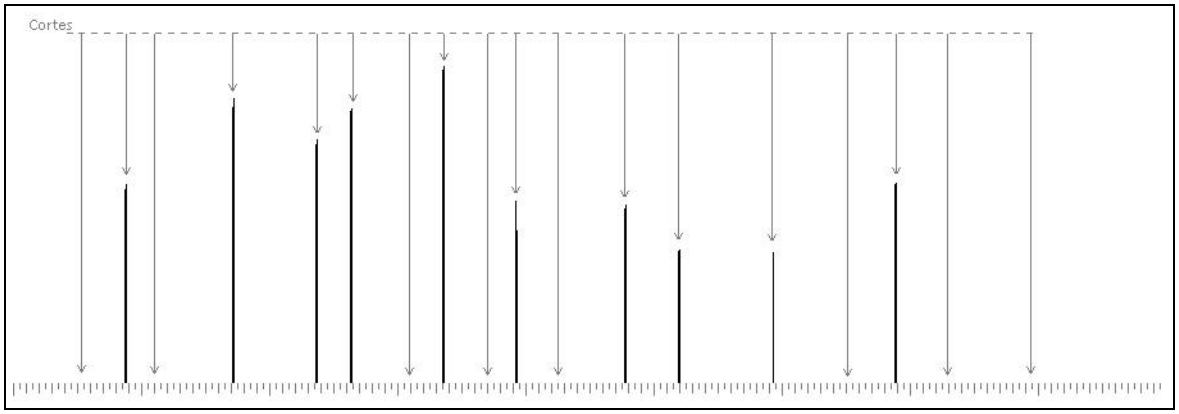
A Figura 3.7 apresenta o resultado da limiarização automática por K-means no resultado da aplicação do WTH_3 apresentado na Figura 3.5. Numa primeira aplicação do K-means, uma diferença muito grande de alguns valores representando cortes pode provocar uma classificação errônea de valores não tão elevados associados a esta mesma classe. A Figura 3.8a ilustra este erro de classificação.

Algoritmo 3.2 K-means duas etapas

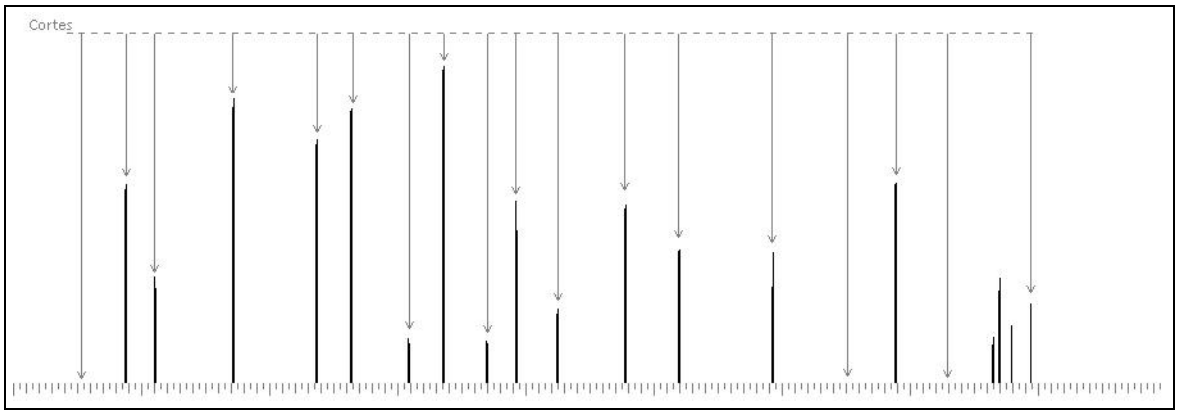
Entrada: vetor de dados V
Saída: dois grupos: $G1$ e $G2$
1: $\langle G1, G2\rangle \leftarrow$ K-means(V);
2: $\langle G1_1, G1_2\rangle \leftarrow$ K-means$(G1)$;
3: $centro1 \leftarrow$ Média dos elementos de $G1_1$;
4: $centro2 \leftarrow$ Média dos elementos de $G1_2$;
5: **if** $centro2 - centro1 > R * centro2$ **then**
6: $\quad G1 \leftarrow G1_1$;
7: $\quad G2 \leftarrow G2 \cup G1_2$;
8: return $\langle G1, G2\rangle$;

A abordagem para a solução deste problema considera uma clusterização adicional do grupo correspondente ao menor centro, após a primeira execução do K-means. O algoritmo proposto (Algoritmo 3.2) realiza primeiramente uma execução do K-means, obtendo dois grupos ou clusters (Algoritmo 3.1). Posteriormente, outra execução do K-means considerando apenas os elementos do grupo de menor centro é realizada, obtendo-se dois novos grupos, $G1_1$ e $G1_2$. Se a distância entre estes for significativa, os elementos do maior grupo, $G1_2$, são considerados como quadros de cortes, caso contrário, estes são

descartados juntamente com os elementos do menor grupo $G1_1$. Considerou-se aqui, como diferença significativa, o valor da constante $R = 0.80$. O resultado dessa aplicação pode ser observado na Figura 3.8b. Os cortes não detectados nesta etapa são todos provenientes de transições em cenários altamente semelhantes. As falsas transições detectadas são devido a mudanças bruscas na cena causadas pela introdução de novos objetos.



a) K-means referente ao Algoritmo 3.1.



b) K-means 2 passos referente ao Algoritmo 3.2.

Figura 3.8: Aplicação do K-means sobre o sinal da Figura 3.6b.

3.2 Detecção de Flashes

O flash pode ser visto como um efeito abrupto caracterizado pela presença de uma alta luminosidade numa seqüência de vídeo, durante um curto intervalo de tempo, em média inferior a um quarto de segundo. A ocorrência de flashes é bastante comum em vídeos de

telejornal, por exemplo, devido a presença de fotógrafos em filmagens referentes a pessoas famosas, artistas, políticos, etc.

As próximas seções apresentam as principais etapas do método proposto para detecção de flashes.

3.2.1 Saturação

O algoritmo de detecção de flashes explora a informação de saturação, presente no VR_{ha}, relativa aos quadros contendo este evento. Esta informação baseia-se no fato de que nas ocorrências de flash, o histograma acumulado do respectivo quadro atingirá um valor máximo acumulado, T_{s2}, a partir de uma posição (*bin*) elevada, T_{s1}. Este vetor pode ser definido como:

Definição 3.7. Saturação do Histograma Acumulado - *Seja o VR_{ha} de um segmento de vídeo digital V de N quadros. Sejam T_{s1} e T_{s2} dois limiares pré-definidos. O vetor $\mathcal{S}$ associado à saturação do histograma acumulado é definido como:*

$$\mathcal{S}(t) = \begin{cases} 1, & \textit{se } \nexists z \in [0, T_{s1}), \textit{tal que } VR_{ha}(t,z) \geq T_{s2} \\ 0, & \textit{caso contrário} \end{cases}$$

em que $t \in [0, N-1]$.

Observe que os limiares T_{s1} e T_{s2} são facilmente definidos em função das características do evento (Figura 3.9), como ilustrado no Capítulo 4.

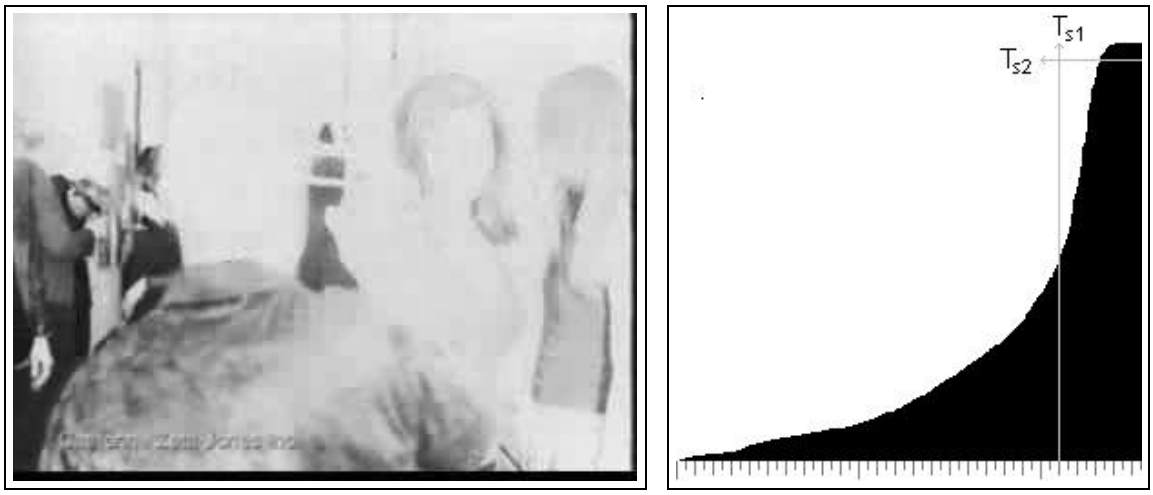
A Figura 3.10a apresenta o vetor de saturação do histograma acumulado para um segmento de vídeo digital com a presença de sete flashes. O mesmo vetor é apresentado na Figura 3.11a, para outro segmento com a ocorrência de um flash.

3.2.2 Intensidade Média

Visando uma maior precisão e filtragem dos quadros candidatos à representação de flashes, a informação de saturação considerada anteriormente pode ser associada ao valor médio da intensidade dos pixels destes quadros.

Durante a ocorrência de um flash, o valor médio dos pixels nos quadros vizinhos tende a aumentar, atingindo um máximo no ponto do flash, e a diminuir rapidamente após o evento. Esta configuração gera picos no relevo do vetor da média das intensidades definido como:

Definição 3.8. Média das Intensidades dos Pixels - *Seja o VR_{ha} de altura H_{VR} e largura N, de um segmento de vídeo digital V de N quadros. O vetor $\mathcal{M}$ de intensidade*



a) Quadro de uma ocorrência de flash. b) Histograma acumulado de a).

Figura 3.9: Comportamento do histograma acumulado para uma ocorrência de flash.

média é dado por:

$$M(t) = \frac{1}{VR_{ha}(t, H_{VR} - 1)} * (VR_{ha}(t, 0) + \sum_{z=1}^{H_{VR}-1} (VR_{ha}(t, z) - VR_{ha}(t, z-1)) * z)$$

$$\mathcal{M}(t) = \begin{cases} M(t), & se\ M(t) >= \frac{L}{2} \\ 0, & caso\ contrário \end{cases}$$

em que $t \in [0, N-1]$*, para* L *níveis de cinza.*

De acordo com as características do flash, os quadros de ocorrência deste evento devem apresentar um alto valor em $\mathcal{M}$. Essa informação é empregada aqui na definição de novos quadros candidatos a flash cujo valor da intensidade média é superior a metade dos níveis de cinza considerados. As Figuras 3.10b e 3.11b apresentam candidatos a flash definidos de acordo com o critério de intensidade média dos quadros.

A seção seguinte combina este dois critérios visando uma melhor validação dos quadros candidatos.

3.2.3 Validação de Candidatos

Após a definição de quadros candidatos baseada nas informações de saturação e intensidade média, uma combinação destes dados pode ser feita visando uma validação destes

candidatos. Esta combinação permite, por exemplo, a remoção de quadros do conjunto de candidatos definidos pela saturação contendo um certo número de pixels de alta intensidade e baixa média global. Estes candidatos não representam, necessariamente, quadros de flash.

A função definida a seguir permite uma combinação das informações anteriores:

Definição 3.9. Vetor de Combinação - *Sejam $\mathcal{S}$ e $\mathcal{M}$ os vetores de saturação do histograma acumulado e de intensidade média, respectivamente, correspondentes a um vídeo digital V de N quadros. O vetor $\mathcal{C}$, constituído da combinação de $\mathcal{S}$ e $\mathcal{M}$, pode ser dado por:*

$$\mathcal{C}(t) = \begin{cases} \mathcal{M}(t), & \textit{se } \mathcal{S}(t) > 0 \\ 0, & \textit{caso contrário} \end{cases} \qquad \forall t \in [0, N-1].$$

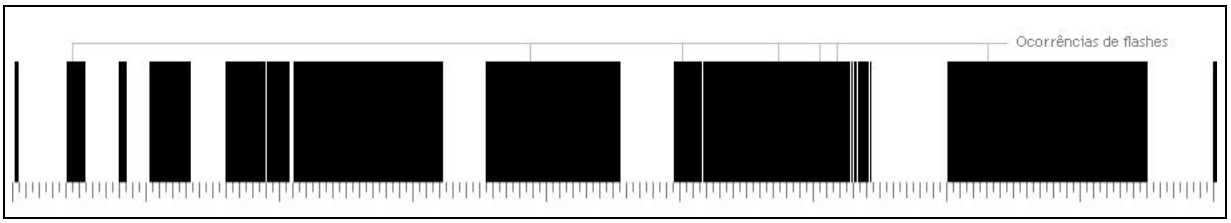
A Figura 3.10c apresenta o vetor $\mathcal{C}$ referente aos vetores das Figuras 3.10a e 3.10b, contendo sete flashes. Observe aqui que o vetor da Figura 3.10b foi validado completamente a partir de 3.10a. Já na Figura 3.11 para a ocorrência de um único flash, o vetor combinação apresentado na c) corresponde a valição completa de a) a partir de b).

3.2.4 Classificação Final

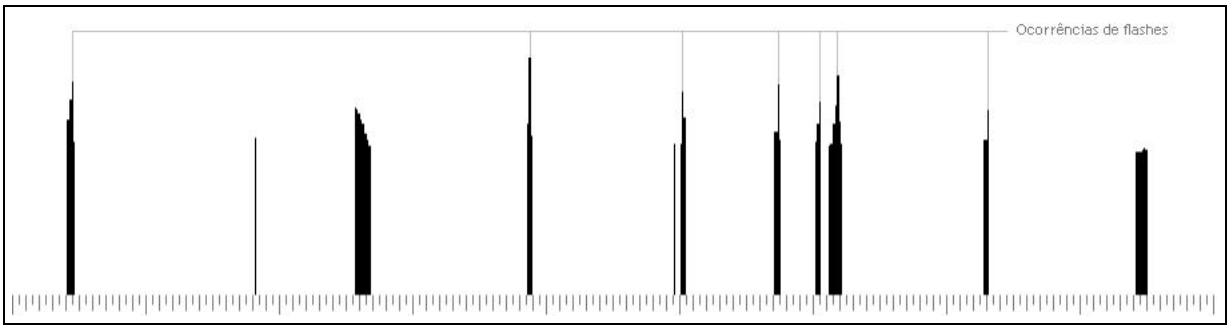
Uma análise do sinal do vetor $\mathcal{C}$ permite a identificação dos principais quadros candidatos a flash. Este conjunto de quadros apresenta um alto valor de intensidade média e saturação, caracterizando picos neste vetor. Para uma filtragem de $\mathcal{C}$ com eliminação de possíveis componentes não representando ocorrências de flash, pode-se considerar uma classificação por K-means ($K = 2$) dos valores diferentes de zero deste sinal. A Figura 3.12 apresenta o resultado desta operação sobre o vetor da Figura 3.10c.

A Figura 3.13 apresenta o vetor $\mathcal{C}$ de um segmento de vídeo sem a ocorrência de flash. Observe que alguns quadros com altos níveis de saturação no VR_{ha} e intensidades médias estão presentes em $\mathcal{C}$, caracterizando uma detecção errônea do evento. A etapa final do método de detecção de flashes proposto aqui consiste da definição dos máximos regionais das componentes conexas do sinal $\mathcal{C}$ (grupos de quadros vizinhos diferentes de zero), e do comprimento (em número de quadros) de cada uma destas componentes.

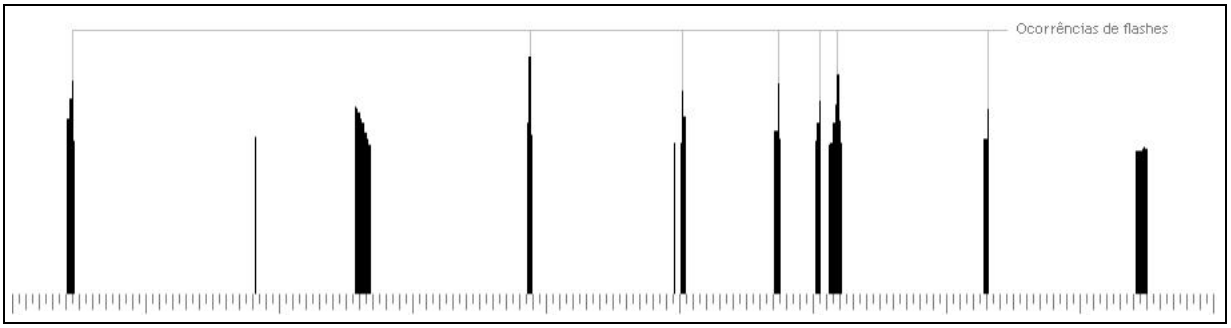
Sabendo-se que os quadros de maior intensidade de um flash tem uma duração média de cinco quadros (para uma freqüência de 29,97 quadros por segundo), as componentes de $\mathcal{C}$ maiores que este limiar podem ser considerados como não representativos deste evento. Os máximos regionais das componentes resultantes, por sua vez, formam o conjunto dos quadros do vídeo contendo a informação procurada.



a) vetor binário da saturação do histograma acumulado.

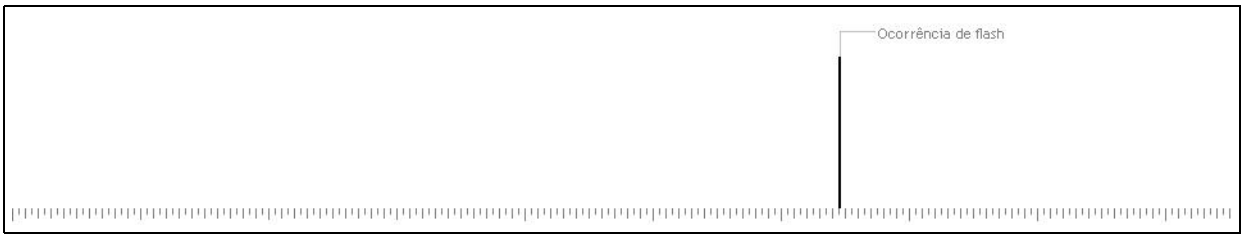


b) vetor das intensidades médias.

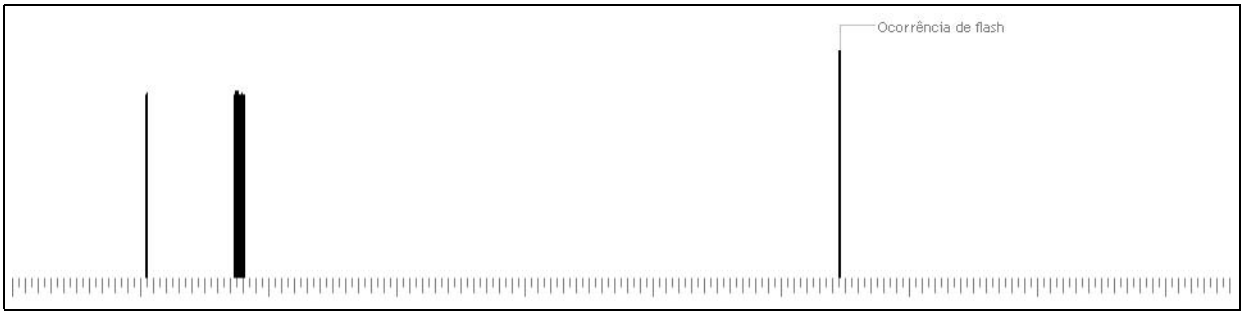


c) vetor de combinação $\mathcal{C}$.

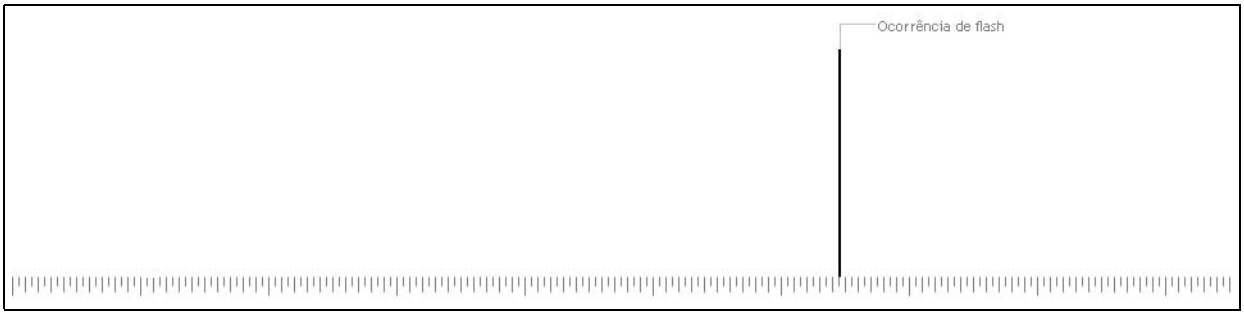
Figura 3.10: Validação de candidatos a flash para um segmento de vídeo.



a) vetor binário da saturação do histograma acumulado.



b) vetor das intensidades médias.



c) vetor de combinação $\mathcal{C}$.

Figura 3.11: Validação de candidatos a flashes para outro segmento de vídeo.

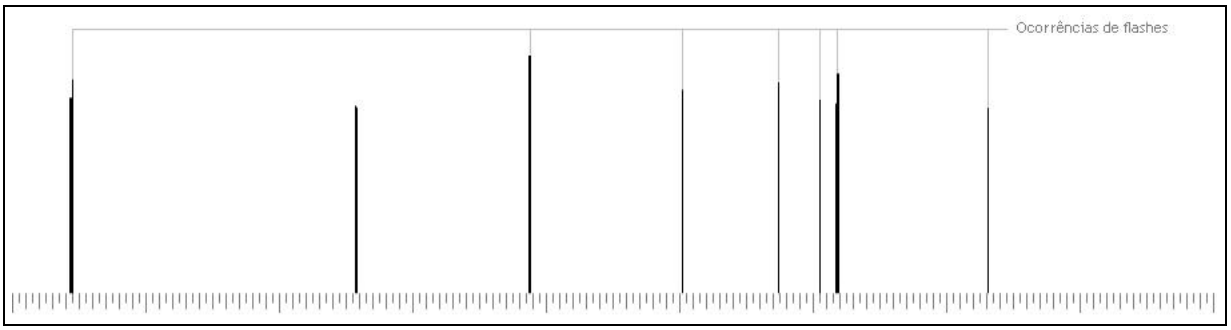
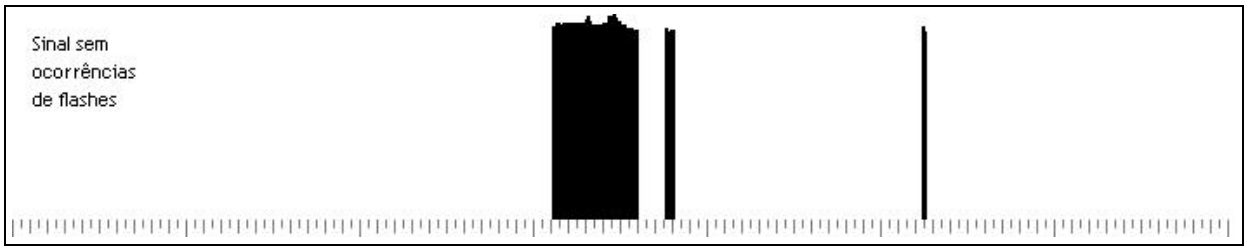
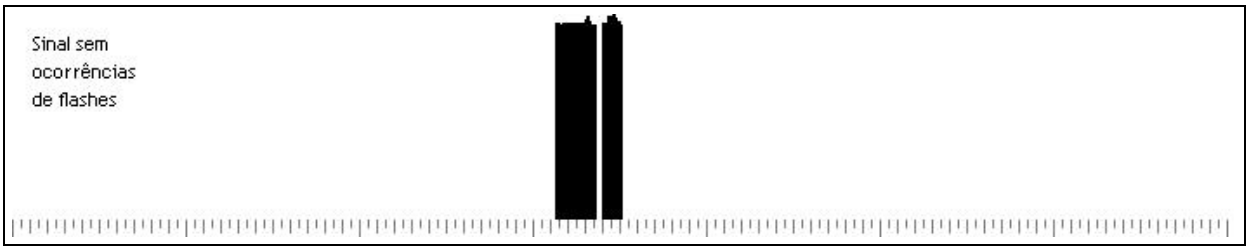


Figura 3.12: Resultado após a classificação do vetor $\mathcal{C}$ da Figura 3.10c.



a) vetor de combinação $\mathcal{C}$.



b) resultado da classificação de a).

Figura 3.13: Classificação para um segmento de vídeo sem ocorrências de flashes.

Capítulo 4

Resultados

Para validação do método proposto, foram realizados alguns testes com vídeos de comerciais de televisão, disponibilizados em sua grande maioria em formato digital pelo site PubsTV [Proom]. Todos os testes foram realizados em um Pentium 4 1.6MHz, com 512 MBytes de memória RAM, no sistema operacional Linux. Serão apresentadas neste capítulo as medidas de qualidade para a avaliação dos principais resultados obtidos.

4.1 Medidas de Qualidades

Em geral, as principais medidas de qualidade, para análise de resultados dos problemas de detecção de tomadas em vídeo digital, estão associadas às detecções que correspondem a um evento procurado, as que não correspondem, e aquelas relativas a eventos não detectados. As Definições 4.1, 4.2 e 4.3, a seguir, definem mais claramente estas medidas.

Definição 4.1. Detecção Correta: *é toda detecção que efetivamente representa um evento previsto de ser detectado pelo algoritmo. O número total de detecções corretas será definido como* N_d.

Definição 4.2. Falsa Detecção (falso positivo): *é qualquer detecção que não representa um evento ou que representa um evento não previsto de ser detectado pelo algoritmo. O número total de falsas detecções será definido como* N_ϕ.

Definição 4.3. Evento não Detectado (falso negativo): *é todo evento presente no vídeo cuja detecção é prevista e que, por um motivo qualquer, não foi detectado pelo algoritmo. O número total de eventos não detectados será definido como* N_μ.

Outras três medidas empregadas na pesquisa e recuperação em bancos de dados de imagens e vídeo também podem ser consideradas aqui. São elas precisão e revocação,

definidas, por exemplo, em [PS97, FT98, Gui03, KTP01], e erro de detecção [Gui03]. Estas medidas são apresentadas a seguir:

Definição 4.4. Precisão (precision): *define a taxa de precisão do algoritmo, conforme a Equação 4.1.*

$$\pi = \begin{cases} 1, & se\ N_d + N_\phi = 0 \\ \frac{N_d}{N_d+N_\phi}, & caso\ contrário \end{cases} \tag{4.1}$$

Definição 4.5. Revocação (recall): *define a taxa de revocação do algoritmo, calculada pela Equação 4.2.*

$$\rho = \begin{cases} 1, & se\ N_d + N_\mu = 0 \\ \frac{N_d}{N_d+N_\mu} & caso\ contrário \end{cases} \tag{4.2}$$

Definição 4.6. Erro de detecção (error): *define a taxa de erro obtida pelo algoritmo, calculada pela Equação 4.3.*

$$\epsilon = \begin{cases} 0, & se\ N_d + N_\mu = 0 \\ \frac{N_\phi}{N_d+N_\mu} & caso\ contrário \end{cases} \tag{4.3}$$

Estas medidas são utilizadas para avaliar os resultados obtidos nesta dissertação, para os métodos de detecção de transições abruptas e flashes.

4.2 Análise dos Resultados

Durante a implementação dos algoritmos, nenhuma sub-amostragem espacial dos quadros foi realizada. Uma subquantização para 128 níveis de cinza foi realizada para o cálculo do histograma acumulado e das intensidades médias. As seções seguintes apresentam os principais resultados obtidos.

4.2.1 Detecção de Flashes

Foram realizados testes com vídeos contendo quadros com alguns flashes e com vídeos sem a presença dos mesmos. De modo geral, os limiares T_{s1} e T_{s2}, apresentados na Definição 3.7, podem ser definidos da seguinte forma:

$T_{s1} = 105$, equivalente a 82% dos bins considerados no histograma acumulado,
$T_{s2} = 95,7\%$ do valor máximo no histograma acumulado.

Os valores apresentados acima foram obtidos a partir da análise dos quadros correspondentes a ocorrências de flashes em alguns dos segmentos de vídeo disponíveis. No entando, esses valores são constantes e independentes do vídeo submetido.

Vídeos	N_d	N_ϕ	N_μ	π	ρ	ϵ
0117	0	0	0	1.0	1.0	0.0
alpharomeo	3	0	0	1.0	1.0	0.0
artikart	10	0	0	1.0	1.0	0.0
aucland	1	0	0	1.0	1.0	0.0
axe	0	0	0	1.0	1.0	0.0
axe-bel	0	0	0	1.0	1.0	0.0
chocol	0	0	0	1.0	1.0	0.0
cocaglacons	0	0	0	1.0	1.0	0.0
doritos	0	0	0	1.0	1.0	0.0
fnac	0	0	0	1.0	1.0	0.0
fu	0	0	0	1.0	1.0	0.0
hyundai	0	0	0	1.0	1.0	0.0
sony	0	0	0	1.0	1.0	0.0
victoria	7	1	0	0.875	1.0	0.125
Total	21	1	0	0.955	1.0	0.048

Tabela 4.1: Resultados da detecção de flashes.

A Tabela 4.1 apresenta os resultados da detecção para um conjunto de vídeos, considerando os parâmetros acima fixados.

Apesar do número ainda reduzido de vídeos disponível para este tipo de teste, estes resultados obtidos indicam um alto grau de discriminação do método entre vídeos com e sem flash e a sua insensibilidade à presença de cortes. De modo geral, o resultado do reconhecimento do evento para este conjunto inicial de testes foi altamente satisfatório.

4.2.2 Detecção de Cortes

Para a detecção de transições do tipo corte, considera-se que todos os vídeos analisados possuem pelo menos uma ocorrência deste evento, juntamente com alguns efeitos de câmera, tais como zooming, panning e flashes (os vídeos com ocorrência de flashes tiveram este evento classificado anteriormente).

Para permitir a comparação do método proposto com outras abordagens conhecidas na literatura, o mesmo conjunto de vídeo foi utilizado considerando as abordagens do teste do χ^2 (Equação 2.6), da diferença de histogramas (Equação 2.5) e do método proposto nesta dissertação. Note que as duas primeiras abordagens utilizam um limiar diferente para cada vídeo. Deve-se ressaltar que para estas abordagens, variou-se os limiares independentemente, permitindo o melhor resultado possível para cada vídeo em cada abordagem.

Vídeos	N_d	N_ϕ	N_μ	π	ρ	ϵ
0117	5	1	0	0.833	1.0	0.200
alpharomeo	17	18	12	0.486	0.586	0.621
artikart	4	57	5	0.066	0.444	6.333
aucland	20	7	5	0.741	0.800	0.280
axe	31	2	0	0.939	1.0	0.065
axe-bel	10	5	2	0.667	0.833	0.417
chocol	9	2	2	0.818	0.818	0.182
cocaglacons	7	1	0	0.875	1.0	0.143
doritos	17	7	7	0.708	0.607	0.250
fnac	10	4	5	0.714	0.667	0.267
fu	25	0	0	1.0	1.0	0.0
hyundai	20	5	2	0.800	0.909	0.227
sony	16	0	2	1.0	0.889	0.0
victoria	29	11	4	0.725	0.879	0.333
Total	220	120	50	0.647	0.815	0.444

Tabela 4.2: Resultados da detecção de cortes utilizando o teste do χ^2.

Vídeos	N_d	N_ϕ	N_μ	π	ρ	ϵ
0117	5	1	0	0.833	1.0	0.200
alpharomeo	21	4	8	0.840	0.724	0.138
artikart	4	16	5	0.200	0.444	1.778
aucland	23	13	2	0.639	0.920	0.520
axe	31	2	0	0.939	1.0	0.065
axe-bel	10	3	2	0.769	0.833	0.250
chocol	10	1	1	0.909	0.909	0.091
cocaglacons	6	1	1	0.857	0.857	0.143
doritos	28	5	0	0.848	1.0	0.179
fnac	15	16	0	0.484	1.0	1.067
fu	25	0	0	1.0	1.0	0.0
hyundai	22	3	0	0.880	1.0	0.136
sony	17	0	1	1.0	0.944	0.0
victoria	29	11	4	0.725	0.879	0.333
Total	246	76	24	0.764	0.911	0.281

Tabela 4.3: Resultados da detecção de cortes utilizando a diferença de histogramas.

Vídeos	N_d	N_ϕ	N_μ	π	ρ	ϵ
0117	5	0	0	1.0	1.0	0.0
alpharomeo	20	1	9	0.952	0.690	0.034
artikart	6	0	3	1.0	0.667	0.0
aucland	21	11	4	0.656	0.840	0.440
axe	23	0	8	1.0	0.742	0.0
axe-bel	9	0	3	1.0	0.750	0.0
chocol	8	1	3	0.889	0.727	0.091
cocaglacons	5	1	2	0.833	0.714	0.143
doritos	22	9	6	0.710	0.786	0.321
fnac	13	2	2	0.867	0.867	0.133
fu	21	0	4	1.0	0.840	0.0
hyundai	18	0	4	1.0	0.818	0.0
sony	15	3	3	0.833	0.833	0.167
victoria	30	6	3	0.833	0.909	0.182
Total	216	34	54	0.864	0.800	0.126

Tabela 4.4: Resultados da aplicação do método para detecção de transições abruptas.

A Tabela 4.2 ilustra os melhores resultados obtidos para as següências de vídeo analisadas. Observou-se que o teste do χ^2 é muito sensível à qualidade do vídeo e a movimentos de objetos. Pode-se observar que a precisão do método foi baixa, e o erro alto, com boa qualidade de revocação.

A Tabela 4.3 exibe os melhores resultados obtidos para as mesmas següências de vídeo. Note que a qualidade do vídeo e e a presença de movimentos de objetos continuou influenciando na precisão do método, embora superior à abordagem anterior, mas ainda relativamente baixa. Por outro lado, a revocação foi bem superior à abordagem anterior.

A Tabela 4.4 apresenta os resultados também obtidos para as mesmas seqüências de vídeo. As principais ocorrências de cortes não detectados deve-se, em parte, ao fato de se utilizar o histograma acumulado como informação global. Verificou-se, por exemplo, que em alguns dos casos de erro as transições associadas possuem quadros com histogramas semelhantes decorrentes, em sua maioria, de um mesmo cenário. Algumas das falsas detecções foram motivadas por bruscas alterações de luminosidades ou introdução de novos objetos, rapidamente acrescentados à cena, provocando intensa variação no histograma acumulado dos respectivos quadros. Ainda assim, comparado aos resultados das abordagens apresentadas anteriormente, a precisão e a revocação foram equivalentemente relevantes. Observe que para este modelo não há a necessidade da intervensão do usuário na definição dos limiares para cada vídeo analisado. Nas duas primeiras abordagens, isso

é fundamental para a qualidade do resultado.

De modo geral, pode-se dizer que os resultados apresentados por este modelo são altamente satisfatórios, comparáveis a outros métodos computacionalmente mais complexos (e caros) definidos na literatura [BMM99, Gui03, NPC99a]. Melhorias no método de detecção de cortes podem ser introduzidas considerando-se, por exemplo, uma partição regular dos quadros e uma posterior definição dos seus histogramas acumulados. A combinação dos resultados destes diferentes VR_{ha}, processados localmente, deve servir a uma caracterização mais completa e precisa do evento.

Capítulo 5

Conclusões

Foi apresentada nesta dissertação uma nova abordagem para detecção de transições e eventos abruptos (cortes e flashes), com algoritmos simples e de baixo custo computacional, considerando-se uma imagem representativa do vídeo digital. Essa imagem, definida aqui como ritmo visual por histograma acumulado, foi proposta para agregar informações globais ao modelo existente do ritmo visual. Os métodos apresentados nesta dissertação permitiram que o problema de detecção de transições em vídeo digital fosse convertido para um problema de detecção de padrões simples em sinal $1\mathbb{D}$.

Baseado essencialmente na informação contida na imagem de ritmo visual por histograma acumulado, a implementação dos dois métodos propostos pôde ser combinada, visando um melhor resultado na detecção dos eventos, refinando-se, por exemplo, a detecção dos cortes.

Em termos de desempenho, os resultados são comparáveis àqueles descritos na literatura para um mesmo conjunto de vídeo [BL03, GLCdAA03b]. Visando uma redução do número de falsos negativos na detecção de cortes, pode-se considerar, por exemplo, um particionamento regular dos quadros e uma posterior definição de seus histogramas acumulados. Uma combinação total ou parcial dos diferentes resultados obtidos para cada VR_{ha}, processados localmente, deve permitir uma caracterização mais completa e precisa do evento através, por exemplo, de diferentes formas de criação da imagem representativa do vídeo.

Observou-se que a subquantização do histograma acumulado para 128 níveis de cinza apresentou resultados superiores àqueles obtidos para 256 níveis. Uma análise mais detalhada desta subquantização deve ser feita, seguida de uma mesma análise referente a alterações na resolução espacial dos quadros.

O método de detecção de transições abruptas pode ser aplicado em seqüências de imagens temporalmente inter-relacionadas como, por exemplo, imagens de satélites, visando uma identificação de mudanças abruptas associadas a eventos típicos, tais como

queimadas e desmatamentos.

Finalmente, conceito de ritmo visual considerado neste trabalho pode ser estendido à identificação de novos eventos e efeitos de câmera de um vídeo digital. Visando uma maior robustez dos algoritmos, métodos estatísticos de detecção e classificação de eventos (classificação bayesiana, lógica fuzzy etc) devem ser incorporados aos mesmos, sem comprometimento das características originais de simplicidade e desempenho computacional. Uma otimização deste desempenho pode ser abordada ainda considerando-se a aplicação do método sobre a representação do sinal no domínio de compressão MPEG.

Referências Bibliográficas

[AG03] S. Arivazhagan e L. Ganesan. Texture segmentation using wavelet transform. *Pattern Recognition Letters*, 24(16):3197–3203, http://dx.doi.org/10.1016/j.patrec.2003.08.005, 2003.

[AM88] J. Aumont e M. Marie. *l'Analyse de films*. Nathan Université, 1988.

[AT94] A. Akutsu e Y. Tonomura. Video tomography: an efficient method for camerawork extraction and motion analysis. Em *Proceedings of the second ACM international conference on Multimedia*, páginas 349–356, http://doi.acm.org/10.1145/192593.192697, 1994.

[BL03] Francisco Nivando Bezerra e Neucimar Jerônimo Leite. Video transition detection using string matching: preliminary results. Em *SIBGRAPI XVI Brazilian Symposium on Computer Graphics and Image Procesing*, páginas 339–346, 2003.

[BM99] Roberto Brunelli e Ornella Mich. On the use of histograms for image retrieval. Em *IEEE International Conference on Multimedia Computing and Systems*, volume 2, páginas 143–147, http://citeseer.nj.nec.com/brunelli99use.html, 1999.

[BMM96] R. Brunelli, O. Mich, e C. Modena. A survey on video indexing. Relatório Técnico 9612-06, Instituto per la Ricerca Scientifica e Tecnologica, http://citeseer.nj.nec.com/brunelli96survey.html, 1996.

[BMM99] R. Brunelli, O. Mich, e C. M. Modena. A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, 10(2):78–112, http://dx.doi.org/10.1006/jvci.1997.0404, 1999.

[Chaml] Daniel Chandler. The 'grammar' of television and film, acessado em dezembro de 2003, http://www.aber.ac.uk/media/Documents/short/gramtv.html.

[CHoQ] Shih-Fu Chang e Peter Humphrey. Videoq: An objected oriented video search engine, acessado em janeiro de 2004, `http://www.ctr.columbia.edu/VideoQ/`.

[CL97] Chueh-Wei Chang e Suh-Yin Lee. Video content representation, indexing, and matching in video information systems. *Journal of Visual Communication and Image Representation*, 8(2):107–120, `http://dx.doi.org/10.1006/jvci.1996.0346`, 1997.

[CLK+99] Min Gyo Chung, Jinho Lee, Hyokman Kim, S. Moon-Ho Song, e WoonKyung M. Kim. Automatic video segmentation based on spatio-temporal features. *Korea Telecom Journal*, 4(1):1–13, `http://eeserver.korea.ac.kr/~bk21/publication/wkkim/wkkim_3.pdf`, 1999.

[CLRS02] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, e Clifford Stein. *Algoritmos: teoria e prática*. Ed. Campus, tradução da 2 edição, 2002. ISBN 85-352-0926-3.

[DB99] Claire Demarty e Serge Beucher. Morphological tools for indexing video documents. Em *IEEE International Conference on Multimedia Computing and Systems*, volume 2, páginas 991–992, `http:// csdl.computer.org/comp/proceedings/icmcs/1999/0253/02/02530991abs.htm`, 1999.

[DLZ00] Mark S. Drew, Ze-Nian Li, e Xiang Zhong. Video dissolve and wipe detection via spatio-temporal images of chromatic histogram differences. Em *IEEE International Conference on Image Processing*, volume 3, páginas 929–932, `http://www.cs.sfu.ca/~li/papers-on-line/icip00-2.pdf`, 2000.

[DRA98] Rakcah Dugad, Krishna Ratakonda, e Narenda Ahuja. Robust video shot change detection. Em *Indian Conference on Computer Vision, Graphics and Image Processing*, páginas 358–364, `http://citeseer.nj.nec.com/50222.html`, 1998.

[FT98] A. Müfit Ferman e A. Murat Tekalp. Efficient filtering and clustering methods for temporal video segmentation and visual summarization. *Journal of Visual Communication and Image Representation*, 9(4):336–351, `http://dx.doi.org/10.1006/jvci.1998.0402`, 1998.

[GCLdAA01] Silvio Jamil Ferzoli Guimarães, Michel Couprie, Neucimar Jerônimo Leite, e Arnaldo de Albuquerque Araújo. A method for cut detection based

on visual rhythm. *SIBGRAPI XIV Brazilian Symposium on Computer Graphics and Image Procesing*, 2001.

[GGml] John Gauch e Susan Gauch. VISION digital video library system, acessado em janeiro de 20004, `http://www.ittc.ukans.edu/~sgauch/DVLS.html`.

[GLCdAA02] Silvio Jamil Ferzoli Guimarães, Neucimar Jerônimo Leite, Michel Couprie, e Arnaldo de Albuquerque Araújo. A directional and parametrized transition detection algorithm based on morphological residues. *SIBGRAPI XV Brazilian Symposium on Computer Graphics and Image Procesing*, 2002.

[GLCdAA03a] Silvio Jamil Ferzoli Guimarães, Neucimar Jerônimo Leite, Michel Couprie, e Arnaldo de Albuquerque Araújo. Identification of video transitions by multi-scale gradient analysis. Relatório Técnico IC-03-13, IC - UNICAMP, `http://www.dcc.unicamp.br/ic-tr-ftp/2003/03-13.ps.gz`, Maio 2003.

[GLCdAA03b] Silvio Jamil Ferzoli Guimarães, Neucimar Jerônimo Leite, Michel Couprie, e Arnaldo de Albuquerque Araújo. Video segmentation based on 2d image analysis. *Pattern Recognition Letters*, 24(7), 2003.

[GSF77] Teofilo Gonzalez, Sartaj Sahni, e W. R. Franta. An efficient algorithm for the kolmogorov-smirnov and lilliefors tests. *ACM Transactions on Mathematical Software*, 3(1):60–64, `http://doi.acm.org/10.1145/355719.355724`, 1977. ISSN:0098-3500.

[Gui03] Silvio Jamil Ferzoli Guimarães. *Video Transition Identification Based on 2D Image Analysis*. Tese de Doutorado, Departamento de Ciência de Computação - Universidade Federal de Minas Gerais, 2003.

[GW92] Rafael C. Gonzalez e Richard E. Woods. *Digital Image Processing*. Addison-Wesley Publishing Company, 1992.

[HN01a] Wei Jyh Heng e King N. Ngan. Automatic shot change detection algorithm using multi-stage clustering for MPEG-compressed videos. *Journal of Visual Communication and Image Representation*, 12(3):364–385, `http://dx.doi.org/10.1006/jvci.2001.0469`, 2001.

[HN01b] Wei Jyh Heng e King N. Ngan. An object-based shot boundary detection using edge tracing and tracking. *Journal of Visual Communication and Image Representation*, 12(3):217–239, `http://dx.doi.org/10.1006/jvci.2001.0457`, 2001.

[JK96] Philippe Joly e Hae-Kwang Kim. Efficient automatic analysis of camera work and microsegmentation of video using spatio-temporal images. *Signal Processing : Image Communication*, 8(4):295–307, 1996.

[Joyns] Robert A. Joyce. A library of video transitions, acessado em janeiro de 2004, `http://www.ee.princeton.edu/~robjoyce/research/transitions/`.

[KC01] Irena Koprinska e Sergio Carrato. Temporal video segmentation. *Signal Processing: Image Communication*, 16(5):477–500, `http://citeseer.nj.nec.com/378900.html`, 2001.

[KLY+01] Hyeokman Kim, Jinho Lee, Jae-Heon Yang, Sanghoon Sull, Woonkyung Michael Kim, e Samuel Moon-Ho Song. Visual rhythm and shot verification. *Multimedia Tools and Applications*, 15(3):227–245, `http://eeserver.korea.ac.kr/~bk21/publication/wkkim/wkkim_3.pdf`, 2001.

[KMN+00] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine Piatko, Ruth Silverman, e Angela Y. Wu. The analysis of a simple k-means clustering algorithm. Em *Proceedings of the Sixteenth Annual Symposium on Computational Geometry*, páginas 100–109, `http://doi.acm.org/10.1145/336154.336189`, 2000. ACM Press. ISBN:1-58113-224-7.

[KTP01] Stelios Krinidis, Sofia Tsekeridou, e Ioannis Pitas. Multimodal interaction for scene boundary detection. *IEEE-Eurasip workshop on Nonlinear Signal and Image Processing*, NSIP01, USA, `http://poseidon.csd.auth.gr/papers/PUBLISHED/CONFERENCE/Krinidis01a/Kri%nidis01a.ps.Z`, 2001.

[LG01] Neucimar Jerônimo Leite e Silvio Jamil Ferzoli Guimarães. Morphological residues and a general framework for image filtering and segmentation. *EURASIP Journal on Applied Signal Processing*, 2001(4):219–229, `http://asp.hindawi.com/volume--2001/S111086570100035X.html`, 2001.

[LYL01] Mee-Sook Lee, Yun-Mo Yang, e Seong-Whan Lee. Automatic video parsing using shot boundary detection and camera operation analysis. *Pattern Recognition*, 34(3):711–719, `http://dx.doi.org/10.1016/S0031--3203(00)00007--8`, 2001.

[MP01] M. K. Mandal e S. Panchanathan. Video segmentation in the wavelet compressed domain. *Journal of Visual Communication and Image Representation*, 12(1):17–28, `http://dx.doi.org/10.1006/jvci.2000.0465`, 2001.

[MPE95] Moving Picture Expert Group - ISO, "Generic Coding of moving pictures and associated audio" - part 2:video, ISO/IEC 13818, 1995.

[NPC98] C. W. Ngo, T. C. Pong, e R. T. Chin. Survey of video parsing and image indexing techniques in compressed domain. *Symposium on Image, Speech, Signal Processing, and Robotics (Workshop on Computer Vision)*, 1:231–236, `http://www.cs.ust.hk/~cwngo/isspr98.html`, 1998.

[NPC99a] C. W. Ngo, T. C. Pong, e R. T. Chin. Camera breaks detection by partitioning of 2D spatio-temporal images in MPEG domain. *IEEE International Conference on Multimedia Computing and Systems*, 1:750–755, `http://www.cs.ust.hk/~cwngo/icmcs99.html`, 1999.

[NPC99b] C. W. Ngo, T. C. Pong, e R. T. Chin. Detection of gradual transitions through temporal slice analysis. *IEEE International Conference on Computer Vision and Pattern Recognition*, 1:36–41, `http://www.cs.ust.hk/~cwngo/cvpr99.html`, 1999.

[NPC00] C. W. Ngo, T. C. Pong, e R. T. Chin. A robust wipe detection algorithm. Em *Asian Conference on Computer Vision*, volume 1, páginas 246–251, `http://www.ifp.uiuc.edu/~cwngo/accv00.html`, 2000.

[NPZ03] Chong-Wah Ngo, Ting-Chuen Pong, e Hong-Jiang Zhang. Motion analysis and segmentation through spatio-temporal slices processing. *IEEE Transactions on Image Processing*, 12(3):341–355, `http://research.microsoft.com/asia/dload_files/group/mcomputing/2003P/i%tip02.pdf`, 2003.

[NT91] Akio Nagasaka e Yuzuru Tanaka. Automatic video indexing and full-video search for object appearances. *Visual Database Systems II*, páginas 113–127, 1991.

[Proom] BBD Production. Pubstv.com, acessado em janeiro de 2004, `http://www.pubstv.com/`.

[PS96] Nilesh V. Patel e Ishwar K. Sethi. Compressed video processing for cut detection. *Visual Image Signal Processing*, 143(5):315–323, 1996.

[PS97] Nilesh V. Patel e Ishwar K. Sethi. Video shot detection and characterization for video databases. *Pattern Recognition*, 30(4):583–592, `http://dx.doi.org/10.1016/S0031--3203(96)00114--8`, 1997.

[PZ99] Greg Pass e Ramin Zabih. Comparing images using joint histograms. *Multimedia Systems*, 7(3):234–240, `"http://citeseer.nj.nec.com/pass99comparing.html`, 1999.

[RHC99] Yong Rui, Thomas S. Huang, e Shih-Fu Chang. Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10(4):39–62, `http://dx.doi.org/10.1006/jvci.1999.0413`, 1999.

[SC96] John R. Smith e Shih-Fu Chang. Tools and techniques for color image retrieval, `http://citeseer.nj.nec.com/133934.html`, 1996.

[Ser82] Jean Serra. *Image Analysis and Matematical Morphology*. Academic Press, 1982.

[SP95] Ishwar K. Sethi e Nilesh Patel. A statistical approach to scene change detection. *SPIE Storage and Retrieval for Image and Video Database III*, 2420:329–338, `http://citeseer.nj.nec.com/sethi95statistical.html`, 1995.

[WCHKdu] Howard Wactlar, Michael Christel, Alex Hauptmann, e Takeo Kanade. Infomedia digital video understanding, acessado em janeiro de 2004, `http://www.informedia.cs.cmu.edu/`.

[WWL98] Min Wu, Wayne Wold, e Bede Liu. An algorithm for wipe detection. Em *IEEE Intlernational Conference on Image Processing*, volume 1, páginas 893–897, `http://citeseer.nj.nec.com/wu98algorithm.html`, 1998.

[Yak76] Yoram Yakimovsky. Boundary and object detection in real world images. *Journal of the ACM*, 23(4):599–618, `http://doi.acm.org/10.1145/321978.321981`, 1976. ISSN:0004-5411.

[ZKS93] HongJiang Zhang, Atreyi Kankanhalli, e Stephen W. Smoliar. Automatic partitioning of full-motion video. *ACM Multimedia Systems*, 1(1):10–28, `http://portal.acm.org/citation.cfm?id=173856.173858&dl=GUIDE&dl=GUIDE`, 1993.

[ZWB+99] W. Zhao, J. Wang, D. Bhat, K. Sakiewicz, e N. Nandhakumar. Improving color based video shot detection. *IEEE International Conference on Multimedia Computing and Systems*, 2:752–756, `http://citeseer.nj.nec.com/271547.html`, 1999.