Segmentação automática de vídeo em cenas baseada em coerência entre tomadas

Tiago Henrique Trojahn

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:_____

Segmentação automática de vídeo em cenas baseada em coerência entre tomadas

Tiago Henrique Trojahn

Orientador: **Prof. Dr. Rudinei Goularte**

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

USP – São Carlos
Abril de 2014

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

T843s

Trojahn, Tiago Henrique
Segmentação automática de vídeo em cenas baseada em coerência entre tomadas / Tiago Henrique Trojahn; orientador Rudinei Goularte. -- São Carlos, 2014.
106 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2014.

1. Segmentação em cenas. 2. Coerência visual. 3. Segmentação de vídeo digital. 4. Personalização e Adaptação de Conteúdo. I. Goularte, Rudinei, orient. II. Título.

Dedicatória

Aos meus pais Neldo e Ingrid por todo o amor, compreensão e dedicação nesta difícil jornada.
Aos meus colegas e amigos que ajudaram direta ou indiretamente na realização deste trabalho.

Agradecimentos

Gostaria de agradecer, inicialmente, aos meus pais por todo o amor, carinho e compreensão durante a realização deste mestrado. Mesmo distantes, sei que posso contar com vocês. Muitíssimo obrigado!

Agradeço aos meus colegas e amigos, pelos diversos momentos de descontração e alegria proporcionados ao longo desses últimos anos: Bruno, Diogo, Flor, João, Johana, Kléberson, Matheus, Olibário, Raíza, Raoni, Kamila, Sandra e Lucas. Que possamos continuar aproveitando as alegrias da vida nos próximos anos!

Um agradecimento especial ao Bruno Lopes pelas muitas conversas e trocas de ideias acerca do trabalho, sem os quais o trabalho poderia não ter sido finalizado. Valeu rapaz!

Agradeço aos atuais e antigos colegas de reuniões semanais, Bruno, Edson, Felipe, Íthalo, Johana, Matheus e Tamires, pelas mais diversas sugestões em apresentações de artigos, prévias e no projeto em si.

Agradeço aos diversos professores e funcionários do instituto, que sempre se dispuseram a ajudar nos momentos de necessidade.

Agradeço ainda a FAPESP e o CNPq pelo apoio financeiro para a realização deste projeto.

Por último, mas não menos importante, agradeço ao meu orientador, Prof. Rudinei Goularte, por toda a ajuda e compreensão ao logo destes últimos três anos, além das diversas sugestões e orientações indispensáveis no desenvolvimento do projeto e desta dissertação em si. Muito obrigado!

Resumo

A popularização de aplicativos e dispositivos capazes de produzir, exibir e editar conteúdos multimídia fez surgir a necessidade de se adaptar, modificar e customizar diferentes tipos de mídia à diferentes necessidades do usuário. Nesse contexto, a área de Personalização e Adaptação de Conteúdo busca desenvolver soluções que atendam a tais necessidades.

Sistemas de personalização, em geral, necessitam conhecer os dados presentes na mídia, surgindo, assim, a necessidade de uma indexação do conteúdo presente na mídia. No caso de vídeo digital, os esforços para a indexação automática utilizam como passo inicial a segmentação de vídeos em unidades de informação menores, como tomadas e cenas. A segmentação em cenas, em especial, é um desafio para pesquisadores graças à enorme variedade entre os vídeos e a própria ausência de um consenso na definição de cena.

Diversas técnicas diferentes para a segmentação em cenas são reportadas na literatura. Uma técnica, em particular, destaca-se pelo baixo custo computacional: a técnica baseada em coerências visual. Utilizando-se operações de histogramas, a técnica objetiva-se a comparar tomadas adjacentes em busca de similaridades que poderiam indicar a presença de uma cena. Para melhorar os resultados obtidos, autores de trabalhos com tal enfoque utilizam-se de outras características, capazes de medir a "quantidade de movimento" das cenas, como os vetores de movimento.

Assim, este trabalho apresenta uma técnica de segmentação de vídeo digital em tomadas e em cenas através da coerência visual e do fluxo óptico. Apresenta-se, ainda, uma série de avaliações de eficácia e de desempenho da técnica ao segmentar em tomadas e em cenas uma base de vídeo do domínio filmes.

Palavras-chave: Segmentação em cenas, coerência visual, segmentação de vídeo digital, Personalização e Adaptação de Conteúdo.

Abstract

The popularization of applications and devices capable of producing, displaying and editing multimedia content did increase the need to adapt, modify and customize different types of media for different user needs. In this context, the area of Personalization and Content Adaptation seeks to develop solutions that meet these needs.

Personalization systems, in general, need to know the data present in the media, thus needing for a media indexing process. In the case of digital video, the efforts for automatic indexing usually involves, as an initial step, to segment videos into smaller information units, such as shots and scenes. The scene segmentation, in particular, is a challenge to researchers due to the huge variety among the videos and the very absence of a consensus on the scene definition.

Several scenes segmentation techniques are reported in the literature. One technique in particular stands out for its low computational cost: those techniques based on visual coherence. By using histograms, the technique compares adjacent shots to find similar shots which may indicate the presence of a scene. To improve the results, some related works uses other features to evaluate the motion dynamics of the scenes using features such as motion vectors.

In this sense, this work presents a digital video segmentation technique for shots and scenes, using visual coherence and optical flow as its features. It also presents a series of evaluation in terms of effectiveness and performance of the technique when segmenting scenes and shots of a custom video database of the film domain.

Keywords: Scene segmentation, visual coherence, digital video segmentation, Personalization and Content Adaptation.

Sumário

Lista de Figuras xii

Lista de Tabelas xiii

Lista de Abreviaturas e Siglas xv

1 Introdução 1
- 1.1 Contexto . . . 1
- 1.2 Motivação . . . 3
- 1.3 Objetivo . . . 5
- 1.4 Organização do texto . . . 5

2 Conceitos Relacionados 7
- 2.1 Vídeo Digital . . . 7
- 2.2 Estrutura Hierárquica de Vídeos . . . 10
- 2.3 Características de Vídeo . . . 12
  - 2.3.1 Histogramas . . . 13
  - 2.3.2 Fluxo Óptico . . . 17
- 2.4 Medidas de Avaliação . . . 18
- 2.5 Técnicas de Segmentação de Vídeo . . . 20
  - 2.5.1 Algoritmos de Seleção de Quadros-Chave . . . 21
  - 2.5.2 Algoritmos de Segmentação em Tomadas . . . 22
  - 2.5.3 Algoritmos de Segmentação em Cenas . . . 25
- 2.6 Considerações finais . . . 29

3 A Técnica de Segmentação 31
- 3.1 Fase Um: Segmentação em Tomadas . . . 32
  - 3.1.1 Extração dos Histogramas . . . 32
  - 3.1.2 Cálculo do Fluxo Óptico . . . 34
  - 3.1.3 Segmentação em Tomadas . . . 36
- 3.2 Fase Dois: Segmentação em Cenas . . . 40
  - 3.2.1 Seleção dos Quadros-Chave . . . 41
  - 3.2.2 Cálculo da Coerência entre Tomadas . . . 42

3.2.3 Criação de Bordas de Cenas 45
3.2.4 Junção de Cenas Similares 46
3.2.5 Junção de Cenas com Dinâmica de Movimento Similar 49
3.2.6 Junção de Cenas Intercaladas por Cenas Dissimilares 52
3.3 Detalhes da Implementação da Técnica 53
3.4 Considerações finais 56

4 Avaliação da Técnica 59
4.1 Descrição da Base Confiável 59
4.2 Testes de Desempenho 63
4.3 Testes de Eficácia 65
4.3.1 Resultados Obtidos para a Segmentação em Tomadas 66
4.3.2 Resultados Obtidos para a Segmentação em Cenas 70
4.3.2.1 Segmentação em Cenas dos Trechos dos Vídeos 71
4.3.2.2 Segmentação em Cenas dos Vídeos Completos 74
4.4 Comparação da Segmentação em Cenas com Trabalhos Relacionados . . . 78
4.5 Limites da Técnica Desenvolvida 81
4.6 Considerações finais 82

5 Conclusões e Trabalhos Futuros 85
5.1 Considerações Iniciais 85
5.2 Resultados Obtidos 85
5.3 Trabalhos Futuros 87

Referências 89

Lista de Figuras

2.1 Ilustração de uma das primeiras tentativas de reproduzir o movimento através de um conjunto de imagens estáticas. Adaptado de (CAIS, 2007). . . . 8
2.2 Uma réplica do zootropo do século XIX, um aparato capaz de apresentar imagens com a ilusão do movimento. Retirado de (Wikipedia, 2013). . . . 8
2.3 Estruturas formadoras de um vídeo, em diferentes níveis de abstração semântica. Adaptado de (Coimbra, 2011). 10
2.4 Exemplo de duas tomadas consecutivas, filmada sobre diferentes ângulos, com uma transição abrupta entre a Figura c) e d). 11
2.5 Três quadros-chave de tomadas adjacentes com alta correlação de cores em tons vermelhos. 12
2.6 Três quadros-chave de tomadas adjacentes com alta correlação de cores entre a primeira e a terceira tomada. 12
2.7 Exemplo de quantização dos 256 níveis da escala de cinza em 4 *bins*. 14
2.8 Exemplo gráfico da normalização de um histograma. 15
2.9 Exemplo gráfico da intersecção entre dois histogramas normalizados. 16
2.10 Exemplo gráfico da distância euclidiana entre dois histogramas normalizados. 17
2.11 Exemplo do cálculo do fluxo óptico entre dois quadros e sua representação gráfica. 18

3.1 Divisão da técnica de segmentação em duas fases: a segmentação em tomadas e a segmentação em cenas. 32
3.2 Fluxograma simplificado descrevendo os procedimentos realizados para a extração dos histogramas dos quadros do vídeo. 33
3.3 Fluxograma simplificado descrevendo os procedimentos realizados para a criação dos vetores de fluxo óptico dos quadros do vídeo. 34
3.4 Fluxograma simplificado descrevendo os procedimentos realizados para a segmentação do vídeo em tomadas. 36
3.5 Exemplo de uma transição de tomadas "evidente", devido à alta dissimilaridade entre os histogramas dos quadros ilustrados em (c) e (d). Nesse exemplo, as figuras (a) até (c) e (d) até (e) formariam duas tomadas distintas. 37
3.6 Exemplo de uma transição gradual do filme "A Beautiful Mind" que não seria detectada devido à alta similaridade dos histogramas entre quadros adjacentes. 39

3.7 Exemplo da divisão dos quadros de entradas entre as janelas deslizantes e seus limiares locais calculados. 39
3.8 Exemplos de uniões de transições para a detecção de transições graduais, onde os quadros em cinza são quadros pertencentes à alguma transição. À esquerda, quadros antes da heurística de transições graduais e, à direita, quadros resultantes após a aplicação da heurística. 40
3.9 Exemplo da seleção dos quadros-chave em um segmento de vídeo com duas tomadas distintas. 41
3.10 Exemplo da diferença entre os valores de BSC e BSCW. 44
3.11 Exemplo de um conjunto de tomadas, que formariam uma cena, mas que são divididas graças ao baixo valor BSCW na tomada ilustrada em (e). . . 46
3.12 Exemplos de falsos positivos que persistem mesmo após a aplicação de métodos para junção de cenas similares e/ou com movimentação similar. . 52
3.13 Arquivo de registro de exemplo para a segmentação em tomadas do arquivo "video.mp4". 55

Lista de Tabelas

3.1 Exemplo de cinco tomadas e seus valores BSCW. 46

4.1 Detalhes dos filmes escolhidos para formar a base confiável. 60
4.2 Detalhes de codificação dos vídeos escolhidos para a base confiável. 61
4.3 Detalhes dos trechos de vídeos extraídos dos filmes que compõe a base confiável. 62
4.4 Número de transições de tomadas e cenas, obtidas na segmentação manual, dos vídeos completos e seus trechos. 63
4.5 Resultados para o teste de desempenho ao executar a implementação da técnica com os trechos da base de vídeos. 65
4.6 Resultados para a segmentação em tomadas dos trechos dos vídeos utilizando três quadros de tolerância. 67
4.7 Resultados para a segmentação em tomada dos vídeos completos utilizando três quadros de tolerância. 67
4.8 Valor F_1 para as dez melhores técnicas reportadas no TRECVid 2005 em relação à segmentação em tomadas. Adaptado de (Smeaton et al., 2010). . 69
4.9 Resultados para a segmentação em cenas dos trechos dos vídeos da base de testes com 120 quadros de tolerância. 71
4.10 Comparação entre os resultados obtidos com ou sem o método de junção baseado no fluxo óptico para a segmentação em cenas dos trechos dos filmes. 73
4.11 Resultados para a segmentação em cenas dos vídeos completos da base de testes com 120 quadros de tolerância. 75
4.12 Comparação entre os resultados obtidos com ou sem o método de junção baseado no fluxo óptico para a segmentação em cenas dos vídeos completos. 77
4.13 Comparação entre a técnica BSC desenvolvida por Rasheed e Shah (2010) e a técnica desenvolvida 79
4.14 Comparação entre a técnica BVW desenvolvida por Chasanis et al. (2009) e a técnica desenvolvida 80

Lista de Abreviaturas e Siglas

BSC	*Backward Shot Coherence*
BSCW	*Backward Shot Coherence Weighted*
BVW	*Bag of Visual Words*
CCH	*Contrast Context Histogram*
CV	*Coerência Visual*
DCT	*Discrete Cosine Transform*
FPS	*Frames per second*
GB	*GigaByte*
Hz	*Hertz*
HD	*High Definition*
HSV	*Hue-Saturation-Value*
ICMC	*Instituto de Ciências Matemáticas e de Computação*
IDE	*Integrated Development Environment*
IMDB	*Internet Movie Database*
MB	*MegaByte*
MSE	*Mean Squared Error*
mm	*Milímetro*
P&A	*Personalização e Adaptação de Conteúdo*
PSNR	*Peak signal-to-noise ratio*
ROC	*Receiver Operating Characteristics*
RGB	*Red-Green-Blue*
SC	*Shot Coherence*
SIFT	*Scale-Invariant Feature Transform*
SBTVD	*Sistema Brasileiro de TV Digital*
SVM	*Support Vector Machine*
TM	*Temporal Memory*
TV	*Televisão*
USP	*Universidade de São Paulo*

Capítulo

1

Introdução

1.1 Contexto

Com o advento e a popularização de uma grande gama de diferentes dispositivos capazes de processar dados multimídia, tais como computadores pessoais, celulares e *tablets*, houve uma mudança no paradigma de produção multimídia, que era produzido majoritariamente por especialistas ou entusiastas. Atualmente, tais dispositivos contam com o auxílio de aplicativos que facilitam a produção e disponibilização de conteúdo por usuários leigos, resultando em um aumento expressivo no volume de informação multimídia disponível. Como consequência, ocorre o problema da "sobrecarga de informação", termo cunhado por Toffler (1984), caracterizado pela dificuldade de se localizar, de modo eficiente, o conteúdo que seja de interesse para o usuário. Uma área recente, no campo da Ciência da Computação, que tenta contribuir com soluções para esse problema é a Personalização e Adaptação de Conteúdo (P&A) (Lu et al., 2011; Manzato, 2011).

A adaptação de conteúdo tem como objetivo disponibilizar material multimídia adequado para cada tipo, condição, estado e conectividade dos mais diversos aparelhos. A personalização, caso particular da adaptação, procura encontrar métodos de customizar e/ou filtrar os dados segundo as preferências, necessidades e interesses de um usuário específico (Magalhães e Pereira, 2004). A personalização de conteúdo é o foco de pesquisa nos últimos anos, através do desenvolvimento de serviços categorizados em seleção de conteúdo, sistemas de recomendação e sistemas de sumarização (Adomavicius e Tuzhilin,

2005). Na seleção de conteúdo, o usuário define, interativamente, critérios de busca de itens multimídia. Na recomendação, itens de possível interesse do usuário são oferecidos automaticamente com base em seu histórico de uso ou em um perfil de preferências. Por fim, a sumarização tem como meta produzir índices daquilo que pode ser relevante ao usuário, oferecendo versões reduzidas do conteúdo, como um trecho significativo de um texto, por exemplo, para que possam ser utilizadas pelos usuários.

Os sistemas de personalização, em geral, apresentam uma necessidade em comum: o conhecimento dos dados contidos no conteúdo, chamados de metadados. Tal necessidade surge na área de P&A para que o sistema possa se adequar às exigências dos usuários. Os metadados podem ser classificados, basicamente, como de baixo ou de alto nível semântico. Metadados de baixo nível semântico descrevem características inerentes à mídia em si, tais como codificação, formato de arquivo, tamanho total, entre outros. Metadados de alto nível, por sua vez, descrevem características conceituais do conteúdo, tais como identificação de pessoas, localidade, assunto, entre outros (Snoek et al., 2005).

A extração de metadados, chamada de Indexação Multimídia, pode ser realizada de modo automático ou manual (Brunelli et al., 1999) e aplicada a diversos tipos de mídia. No caso particular de vídeo digital, o primeiro passo é o de segmentar o vídeo em unidades menores de informação, mais gerenciáveis, como quadros, tomadas e cenas, para então aplicar a extração de metadados (Chaisorn et al., 2003; Chen e Li, 2010; Liu et al., 2009; Ogawa et al., 2008; Wang et al., 2008a; Yu et al., 2007a).

A segmentação de vídeo em quadros é uma área já estabelecida, com a existência de técnicas maduras tanto para vídeos comprimidos como para vídeos sem compressão. A área de segmentação de vídeo em tomadas ainda é uma área de pesquisa ativa (Fu e Zeng, 2009; Hong-cai et al., 2010; Manjunath et al., 2011; Zhang e Wang, 2012), que procura superar desafios como a dependência de domínio e limitações quanto a efeitos de transição complexos. Segundo Hanjalic (2002), a segmentação do vídeo em tomadas é especialmente importante, pois segmentações com maior grau de abstração dependem dela. A segmentação em cenas tem atraído a atenção de diversos trabalhos recentes (Baber et al., 2013; Kumar et al., 2011; Sawai et al., 2011; Song et al., 2010) e apresenta ainda muitos desafios. Esse tipo de segmentação é importante devido ao fato de "cena" ser um conceito subjetivo bem difundido entre usuários comuns. Contudo, tal subjetividade impõe dificuldades como a falta de uma definição consensual, resultando em uma área ainda pouco investigada e com grande variação entre resultados obtidos. Segmentações desse tipo potencializam a utilização de aplicações voltadas ao público em geral, seja em anotações, seja na criação de resumos de um vídeo ou na localização de cenas específicas em uma base de vídeos de larga escala, auxiliando a reduzir a sobrecarga de informação.

1.2 Motivação

O grupo de Sistemas Web e Multimídia Interativos do ICMC-USP têm como uma de suas linhas de atuação a área de P&A (Cattelan et al., 2008; Manzato et al., 2011, 2009a; Manzato e Goularte, 2010, 2012; Manzato et al., 2009b), com especial atenção a vídeo, a qual depende do suporte de segmentação. O grupo, contudo, não possui ferramentas automáticas de segmentação adequadas às suas necessidades. Por exemplo, Manzato (2011) desenvolveu um *framework* de personalização e recomendação baseado em metadados semânticos do vídeo. Seu funcionamento consiste na segmentação automática de vídeos em tomadas, através de algoritmos genéticos (Manzato e Goularte, 2007) seguido por uma segmentação manual em cenas. Apesar de a precisão da segmentação em tomadas ser razoável, o uso de algoritmos genéticos eleva o custo computacional (Manzato e Goularte, 2007). Além disso, o *framework* exige um especialista para a segmentação em cenas, em um processo cansativo e lento. Para superar tal limitação, uma técnica automática para a segmentação em cenas poderia ser adotada.

Nesse sentido, a literatura reporta uma ampla gama de técnicas para a segmentação de cenas. As técnicas em estado-da-arte, porém, podem ser agrupadas em duas abordagens: baseadas em aprendizagem de máquina ou baseadas em multimodalidade. Técnicas baseadas em aprendizagem de máquina utiliza-se de métodos tais como algoritmos genéticos ou redes neurais para modelar, identificar, agrupar e classificar a estrutura do vídeo, encontrando, assim, as cenas. Apesar dos bons resultados, como taxas de acurácia próximas a 80% (Wu et al., 2011), técnicas baseadas em aprendizagem de máquina necessitam de treinamento prévio e podem fazer uso de funções de *fitness* complexas, tornando tais técnicas caras computacionalmente.

Técnicas multimodais, por sua vez, realizam a segmentação utilizando mais de um tipo de mídia. Por exemplo, tais técnicas podem realizar a segmentação no vídeo e no áudio e, posteriormente, fundir os resultados em busca de uma segmentação única. O objetivo, nesse tipo de técnica, é o de que segmentações em mais de um fluxo possam compensar problemas na segmentação de fluxos únicos, resultando em uma segmentação mais consistente e, possivelmente, com maior nível semântico. Infelizmente, assim como nas técnicas baseadas em aprendizagem de máquina, o custo computacional também é elevado (Coimbra e Goularte, 2009; Manzato et al., 2011).

Outra técnica de segmentação em cenas, menos recente, é baseada em coerência visual (CV) (Kender e Yeo, 1998). Nesse tipo de técnica, a "coerência" de alguma característica (do Inglês *feature*) entre diferentes segmentos de vídeo é medido com o objetivo de formarem segmentos com alta similaridade visual, como cenas. Porém, paralelamente, enquanto tais técnicas estavam sendo desenvolvidas, as técnicas multimodais e as basea-

das em aprendizagem de máquina começaram a apresentar resultados mais significativos, fazendo com que as técnicas baseadas em CV ficassem em segundo plano.

Uma grande vantagem de técnicas baseadas em CV é que são computacionalmente simples e baratas, devido ao fato de utilizarem características bem conhecidas e de fácil obtenção, como histogramas (Chen e Li, 2010; Rasheed e Shah, 2003). Além disso, tais técnicas podem ser significativamente melhoradas de forma a obter resultados comparáveis com as técnicas multimodais e baseadas em aprendizagem de máquina. Ainda, cabe destacar que o uso de técnicas baseadas em CV não excluí sua utilização conjunta com outras técnicas, como as multimodais.

Originalmente, as técnicas baseadas em CV faziam uso apenas de características de cores para estimar a coerência entre dois segmentos quaisquer (Kender e Yeo, 1998). Para isso, adotavam os histogramas, uma característica bem conhecida e com baixo custo computacional. Com os histogramas dos segmentos estabelecidos, uma medida de dissimilaridade entre os histogramas de dois segmentos é definida e, caso sejam considerados similares, são unidos em um novo segmento. Outros trabalhos (Chen e Li, 2010; Rasheed e Shah, 2003) agregam uma outra característica à análise da CV: o "movimento" presente em cada segmento. Assim, além da análise das cores, a dinâmica de movimento foi analisada como forma de aprimorar os resultados, principalmente em segmentos adjacentes com baixa similaridade de cores, mas com alta similaridade da dinâmica de movimento.

Dentre essas técnicas, destaca-se a técnica desenvolvida por Rasheed e Shah (2003), que propôs uma análise da dinâmica de movimento através da análise dos vetores de movimento presentes no fluxo comprimido do vídeo de entrada. Nesse trabalho, os autores destacam a baixa eficácia da análise de CV baseada em histogramas em cenas com alta movimentação, como em cenas de ação. Assim, os autores estimam a quantidade de movimento e são capazes de unir segmentos adjacentes com alta movimentação, reduzindo o número de falsos positivos obtidos na análise baseada em histogramas de cor.

Infelizmente, mesmo com a adoção da estimação da dinâmica de movimento, as técnicas baseadas em CV apresentam duas desvantagens importantes:

- Baixa precisão (Chen e Li, 2010), principalmente se comparadas às técnicas baseadas em aprendizagem de máquina.

- Dependência do formato de vídeo de entrada (Chen e Li, 2010; Rasheed e Shah, 2003).

Assim, abre-se a possibilidade de desenvolver uma técnica de segmentação em cenas baseada em CV que seja independente do formato de vídeo de entrada e que, ao mesmo tempo, seja capaz de produzir resultados mais significativos, contribuindo para a área de segmentação de vídeo.

1.3 Objetivo

Este trabalho tem como objetivo desenvolver uma técnica automática de segmentação de vídeo em cenas, baseada em coerência visual, que seja independente do formato de vídeo de entrada. A técnica deve fazer uso apenas de características que possam ser extraídas ou calculadas diretamente do fluxo de vídeo, alcançar eficiência comparável a técnicas similares tidas como estado da arte e utilizar menos recursos computacionais.

1.4 Organização do texto

Esta dissertação está organizada em quatro capítulos, além desta introdução. O Capítulo 2 apresenta, para a correta compreensão deste trabalho, uma série de conceitos fundamentais da área de segmentação de vídeo, tais como: conceitos de vídeo, tomadas e cenas; características de vídeo, como histogramas e o fluxo óptico; medidas de avaliação entre técnicas de segmentação de vídeo e uma discussão sobre trabalhos relacionados. Nesse último ponto, em particular, é feita uma análise crítica de trabalhos encontrados na literatura, destacando o estado da arte e justificando as escolhas realizadas para conduzir os testes comparativos.

O terceiro capítulo apresenta a técnica de segmentação em cenas desenvolvida, a qual possui duas fases. Na primeira ocorre a segmentação em tomadas e a extração de informações importantes (como o cálculo de histogramas e do fluxo óptico) para a etapa seguinte; a segunda etapa utiliza as informações obtidas anteriormente para calcular a coerência entre tomadas adjacentes e, assim, detectar os limites entre cenas.

O quarto capítulo apresenta os testes realizados com o objetivo de avaliar, em termos do desempenho e da eficácia, a técnica desenvolvida. Descreve-se: como a base confiável para realizar os testes foi obtida; a metodologia e as métricas utilizadas; o objetivo e o desenvolvimento de cada teste; os resultados obtidos.

Por fim, no quinto e último capítulo, são apresentadas as conclusões e possíveis trabalhos futuros.

Capítulo

2

Conceitos Relacionados

Neste capítulo serão apresentados os conceitos fundamentais para o entendimento deste trabalho. A Seção 2.1 apresenta a definição e algumas características de vídeos digitais. A Seção 2.2 discute a definição de tomadas e cenas utilizadas neste trabalho. Já a Seção 2.3 apresenta duas características de vídeo exploradas pela técnica de segmentação de vídeo: os histogramas e o fluxo óptico.

A Seção 2.5 discute algumas técnicas bem conhecidas e outras no estado-da-arte para a seleção de quadros-chave, e a segmentação em tomadas e em cenas. Por fim, a Seção 2.6 apresenta algumas considerações finais quanto a este capítulo.

2.1 Vídeo Digital

Vídeo, do latim *videre* (ver), pode ser definido como um "conjunto de imagens estáticas que transmitem a sensação de movimento" (Chapman e Chapman, 2009). Um vídeo, então, é formado por um conjunto de imagens estáticas, chamadas de quadros ou *frames*, que são apresentadas a uma taxa constante de tal modo que dão a ilusão de movimento.

Max Wertheimer descreveu, em 1912, o fenômeno conhecido como "movimento beta", onde um conjunto de imagens apresentadas sequencialmente dá a ilusão que os objetos representados movem-se (Sarris, 1987). Segundo Chapman e Chapman (2009), a percepção de movimento dá-se à taxa de cerca de 40 Hz. Tal valor depende de diversos fatores tais como luminosidade, brilho e contraste, tanto do local quanto dos quadros em si.

A ideia de representar o movimento como uma série de imagens com poucas variações entre elas não é nova. Uma tigela de barro de mais de cinco mil anos encontrada em *Shahr-i Sokhta*, no Irã, contém cinco imagens que representariam uma cabra pulando para obter comida em um ramo de folhas. A Figura 2.1 ilustra as imagens encontradas.

Figura 2.1: Ilustração de uma das primeiras tentativas de reproduzir o movimento através de um conjunto de imagens estáticas. Adaptado de (CAIS, 2007).

Já no século XIX, diversos aparatos foram desenvolvidos com o objetivo de apresentar ao público um conjunto de imagens com a ilusão do movimento. O Zootropo, inventado em 1833 pelo britânico William George Horner (Kristin Thompson, 2009), é um exemplo entre diversos outros mecanismos inventados. A Figura 2.2 ilustra uma réplica do zootropo do século XIX.

Figura 2.2: Uma réplica do zootropo do século XIX, um aparato capaz de apresentar imagens com a ilusão do movimento. Retirado de (Wikipedia, 2013).

Em termos cinematográficos, a produção de vídeo baseava-se na filmagem consecutiva de quadros em um filme de determinada largura, como 35 mm. O filme, uma película formada por milhões de cristais de Halogeneto de Prata, ao ser tocado por uma fonte de luz, sofre um processo de reordenação dos cristais, formando imagens. A introdução

da cor no processo de gravação, anos mais tarde, resultou em filmes com três camadas separadas: uma para o vermelho, outra para o verde e a última para o azul.

Com o aumento do poder computacional e a popularização do computador pessoal, a partir da década de 1980, nasceu o desejo de armazenar e assistir vídeos em um computador. Surge então o vídeo digital, onde o conjunto de imagens formadores do vídeo é armazenado em um fluxo de dados. O processo de conversão de imagens analógicas armazenadas em um filme para um vídeo digital é realizado através da "leitura" de cada imagem do filme, resultando em um arquivo contendo os dados de todas as imagens lidas.

Inicialmente, a conversão analógica para digital era dificultada pelo volume massivo de informações e pela escassez ou até inexistência de métodos de codificação capazes de reduzir o tamanho resultante do fluxo de vídeo. Recentemente, o avanço na eficiência de codificadores de vídeo e o crescimento da capacidade computacional tornaram possível o armazenamento de várias horas de vídeo em um dispositivo portátil, como um *tablet*.

Um vídeo digital, assim como o analógico, apresenta uma série de características, algumas das quais estão apresentadas a seguir (Richardson, 2010):

- **Número de quadros por segundo:** um vídeo possui uma determinada taxa na qual os quadros são apresentados ao espectador. A taxa varia conforme diferentes padronizações entre países, modo de armazenamento, entre outros. Utiliza-se, usualmente, a forma abreviada fps, do Inglês *frames per second*, para se referir a essa taxa.

- **Resolução:** a resolução é o número de *pixels* na horizontal e na vertical. Um vídeo é dito de resolução A x B quando seus quadros possuírem A *pixels* na horizontal e B *pixels* na vertical.

- **Colorimetria:** assim como imagens, um dos detalhes mais importantes de um vídeo é a sua colorimetria. Vídeos e imagens podem ser de diferentes espaços de cores, como o *Red-Green-Blue* (RGB) ou o YCbCr, entre outros.

- **Detalhes de codificação:** um vídeo é geralmente codificado para que possa ser armazenado em um espaço de armazenamento tangível. Ao codificar um vídeo, surgem variáveis como o codificador empregado, taxa de compressão, taxa de *bits* por segundo, relação PSNR (Salomon et al., 2006), quantização, vetores de movimento, entre outros.

Algoritmos de segmentação de vídeo sofrem interferência direta das características supracitadas. Por exemplo, vídeos com maior taxa de quadros por segundo e/ou maior resolução exigem maior poder de processamento, devido ao maior volume de informações presentes em um determinado espaço de tempo. Além de características da mídia em

si, um vídeo possui conteúdos com alguma semântica embutida, assim, na Seção 2.2, apresenta-se uma divisão quanto ao conteúdo do vídeo em diferentes níveis de abstração.

2.2 Estrutura Hierárquica de Vídeos

Um vídeo, como apresentado na Seção 2.1, é formado por um conjunto de imagens (quadros ou *frames*) que são apresentadas a uma determinada taxa. Os quadros consecutivos podem ser agrupados em uma ou mais tomadas. Por fim, um conjunto de tomadas consecutivas pode ser agrupado formando uma cena. A Figura 2.3 ilustra as estruturas de um vídeo em diferentes níveis de abstração, partindo do quadro, passando pela tomada e chegando à cena.

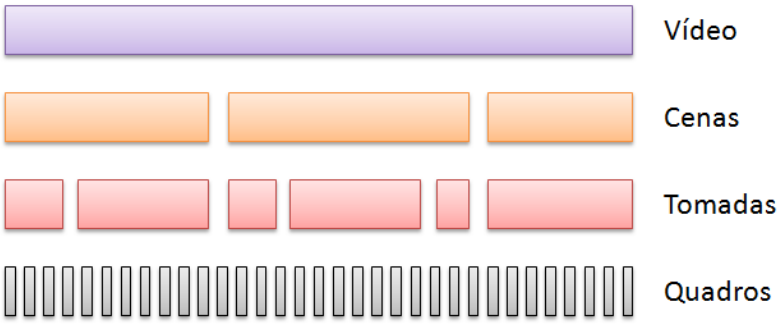


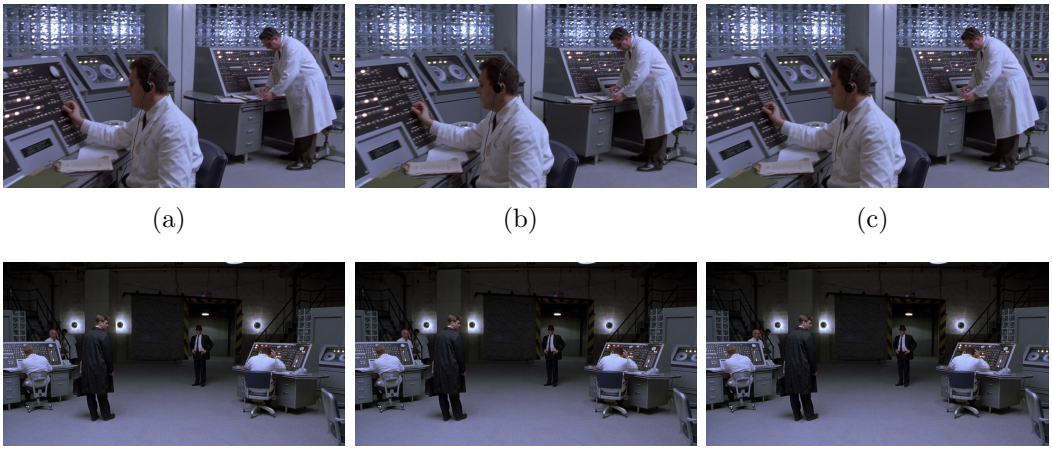
Figura 2.3: Estruturas formadoras de um vídeo, em diferentes níveis de abstração semântica. Adaptado de (Coimbra, 2011).

Uma tomada pode ser formalmente definida como um conjunto de imagens estáticas obtidas continuamente por uma única câmera (Koprinska e Carrato, 2001), representando uma ação contínua no tempo e espaço (Zhang et al., 2007). A Figura 2.4 exemplifica uma tomada.

A interrupção de uma tomada demarca o fim de uma e o começo da próxima tomada, determinando a chamada "transição de tomadas". As transições de tomadas podem ser classificadas em transições abruptas ou graduais, cujas diferenças são descritas a seguir.

- **Transição Abrupta:** quando há uma mudança significativa entre dois quadros adjacentes, podendo ou não representar a troca do conteúdo semântico (por exemplo, demonstrar o mesmo cenário em diferentes ângulos).

- **Transição Gradual:** quando alguns quadros de ambas as tomadas adjacentes são utilizadas para "suavizar" a transição entre as tomadas. São de três tipos básicos (Koprinska e Carrato, 2001): dissolução, onde a imagem antiga é dissolvida em outra imagem *fade in* e *fade out*, quando uma imagem é clareada ou escurecida



(a) (b) (c)

(d) (e) (f)

Figura 2.4: Exemplo de duas tomadas consecutivas, filmada sobre diferentes ângulos, com uma transição abrupta entre a Figura c) e d).

gradativamente e *wipe*, quando uma imagem dá lugar à outra imagem através de um "empurrão".

Embora tomada seja um conceito bem definido, o mesmo não ocorre com cena. Por se tratar de um conceito subjetivo, diversas definições são apresentadas pela literatura sendo que, em geral, cada trabalho na área de segmentação em cenas apresenta uma definição própria.

Algumas definições consideram cenas, por exemplo, como "unidades de informação" (Altheide, 1985) e "cenas visuais" (Graber, 1990). A primeira definição, entretanto, apresenta como desvantagem o fato de ser muito genérica, resultando em dificuldades para definir as transições de cenas. Do mesmo modo, a segunda definição também é excessivamente genérica, impondo dificuldades para determinar a transição de cenas.

Outros trabalhos definem cena como um conjunto de uma ou mais tomadas semanticamente correlacionadas (Huang et al., 1998). Tal definição é menos genérica que as anteriores, pois se utiliza do conceito bem definido de tomada, sendo que o correlacionamento semântico pode ser encontrado por técnicas utilizando correlação entre tomadas (Chen e Li, 2010; Rasheed e Shah, 2003), dicionários visuais (Chasanis et al., 2009), entre outros.

Neste trabalho, foi adotada a definição apresentada por Huang et al. (1998), de modo que:

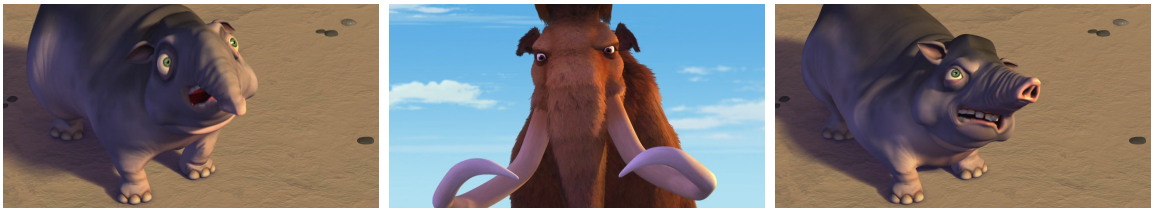
- Uma cena é composta por um conjunto finito e não nulo de tomadas consecutivas.
- Uma transição de cena coincide com a transição de tomadas, ou seja, não ocorre a transição de cena dentro de uma mesma tomada.

- Uma cena contém relacionamento semântico, que, no caso deste trabalho, é dado pela CV e/ou pela dinâmica de movimento entre as tomadas.

A Figura 2.5 ilustra três quadros-chave de três tomadas diferentes adjacentes, onde, embora apresentem variações nos personagens apresentados em cada tomada, a semântica da cena é similar, sendo que a correlação entre as tomadas pode ser visualizada pelas cores e a movimentação similar. Já a Figura 2.6 ilustra três quadros-chave de três tomadas diferentes adjacentes, cuja semântica e o movimento entre as tomadas também é similar, porém com baixa correlação entre as cores da primeira para a segunda tomada e da segunda para a terceira tomada.

(a) Tomada 1 (b) Tomada 2 (c) Tomada 3

Figura 2.5: Três quadros-chave de tomadas adjacentes com alta correlação de cores em tons vermelhos.



(a) Tomada 1 (b) Tomada 2 (c) Tomada 3

Figura 2.6: Três quadros-chave de tomadas adjacentes com alta correlação de cores entre a primeira e a terceira tomada.

É possível notar que, dessa forma, o conceito de cena engloba casos onde tomadas adjacentes são visualmente dissimilares, mas que são relacionadas semanticamente.

2.3 Características de Vídeo

As características que podem ser extraídas de um vídeo são a base da recuperação e da indexação de vídeo (Hu et al., 2011). Segundo Hu et al. (2011), as características visuais podem ser: estáticas, orientada a objetos ou de movimento:

- Características estáticas, como a cor, textura e a forma de objetos são obtidas através de uma análise quadro-a-quadro.

- Características baseadas em objeto são aquelas baseadas em objetos específicos apresentados na tela. Uma importante aplicação é a de extração de faces para encontrar pessoas específicas em um vídeo (Sivic et al., 2005). Embora interessante, tais técnica limitam-se a encontrar apenas alguns objetos pré-definidos, tornando o processo computacionalmente caro e limitado.

- Características baseadas em movimento são aquelas apresentadas tanto pelo movimento da câmera como dos objetos em si tais como quantidade de movimento, direção de deslocamento, tempo para sair da imagem, entre outros. Tais características tem grau diferenciado de importância dependendo do método utilizado.

A literatura reporta inúmeras características presentes em vídeos, visuais e não visuais, em fluxos comprimidos ou não. Cita-se, como exemplo, o ritmo visual (Kim et al., 2001), os blocos de *pixels*, a presença de textos (Yuan et al., 2011), faces (Viola e Jones, 2001), sombras (Wang, 2007), silêncio em fluxos de áudio (Coimbra e Goularte, 2009), *wavelets* (Devasena e Hemalatha, 2012), cores, movimento e bordas de imagens (Lefèvre et al., 2003), entre outros.

Nesta seção são apresentadas as duas características utilizadas na técnica desenvolvida para a segmentação em tomadas e cenas: os histogramas e o fluxo óptico. Os histogramas, uma característica estática, foram escolhidos devido ao fato de representarem características de cor do vídeo, uma das principais informações visuais disponíveis, além de ter um custo computacional baixo. Já o fluxo óptico, uma característica baseada em movimento, foi escolhido pois é um método robusto e independente de formato de vídeo, sendo capaz de identificar a movimentação de objetos e de vídeos em um vídeo. Os histogramas são apresentados na Subseção 2.3.1. Já o fluxo óptico, é descrito na Subseção 2.3.2.

2.3.1 Histogramas

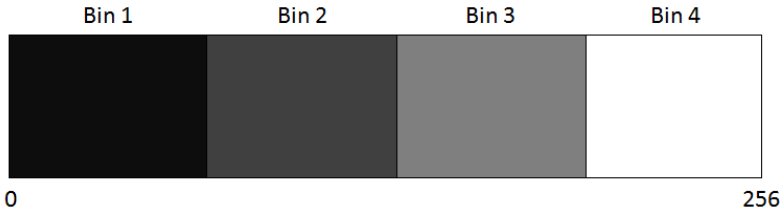
Segundo Marques (2011), o histograma de uma imagem é uma representação gráfica da frequência de cada nível de cor em uma imagem. A importância dos histogramas reside no fato de ser uma representação com volume de dados reduzido de uma imagem, tornando o processamento mais rápido.

Histogramas podem ser calculados sobre vídeos em escala-de-cinza ou coloridos, através da decomposição dos canais de cor (Marques Filho e Vieira Neto, 1999). Caso o vídeo seja em escala-de-cinza, haverá 256 valores possíveis para os *pixels*, podendo-se, então, criar um histograma com 256 níveis. Porém, caso o vídeo seja de algum espaço de cor

como o RGB ou HSV (*Hue-Saturation-Value*), o histograma resultante teria quase 17 milhões de possíveis cores a serem representadas (considerando todas as combinações entre cada nível de cor), 256 para cada canal de cor.

Assim, para reduzir tal volume de dados, o vídeo colorido sofre o processo de quantização (Marques, 2011): cores próximas são agrupadas em um conjunto e, dentro desse conjunto, todas as cores são igualadas. Ao número de grupos da quantização costuma-se dar o nome de *bins* (caixas). A Figura 2.7 ilustra o processo de quantização com 4 *bins* do cinza.

(a) A escala-de-cinza dividida em 256 níveis.



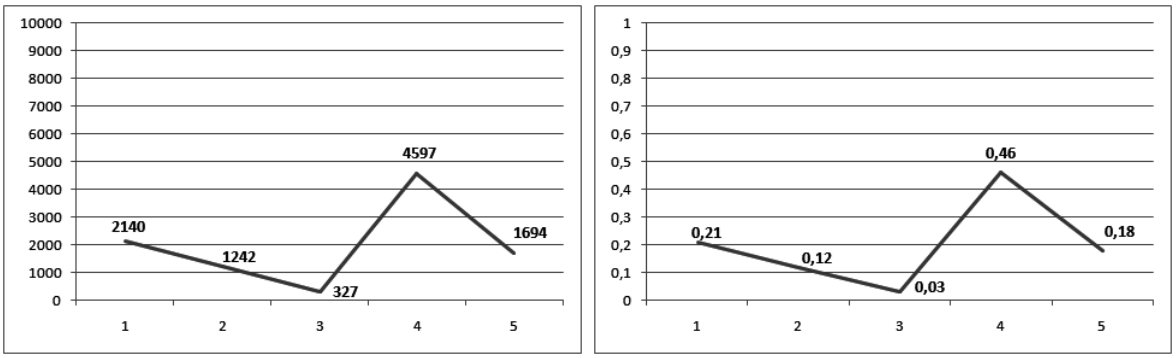
(b) A escala-de-cinza original dividida em 4 *bins*.

Figura 2.7: Exemplo de quantização dos 256 níveis da escala de cinza em 4 *bins*.

Assim, um vídeo no espaço de cor RGB com 12 *bins* na proporção 4:4:4 (4 para vermelho, 4 para verde e 4 para azul) seria representado em apenas 64 valores contra quase 17 milhões de valores sem o processo de quantização. Esse valor é obtido através da combinação de todos os quatro valores de vermelho, quatro para verde e quatro para azul, formando então 64 combinações entre elas. Esse processo apresenta vantagens tanto na redução do tamanho do histograma como em velocidade de processamento de operações sobre o histograma resultante. Como desvantagem, cita-se que a divisão de um histograma em *bins* resulta na perda irreversível de informações tais como pequenos detalhes da imagem, principalmente em degrades.

Um detalhe importante diz respeito aos valores em cada nível do histograma: caso sejam comparados dois histogramas de imagens de resoluções diferentes, o resultado obtido poderá ser incorreto. Tal erro reside do fato de que a quantidade de *pixels* em cada imagem (e, consequentemente, em seus histogramas) é diferente. Assim, para solucionar tal problema, geralmente é aplicado um processo de normalização dos valores. Neste processo, o somatório de todos os *pixels* do histograma é considerado como 1.0 (ou seja,

100% da imagem). A partir disso, todos os níveis de cores do histograma terão seu valor de proporção referente ao número total de *pixels*. Um exemplo da normalização é ilustrado na Figura 2.8



(a) Histograma original. (b) Histograma normalizado.

Figura 2.8: Exemplo gráfico da normalização de um histograma.

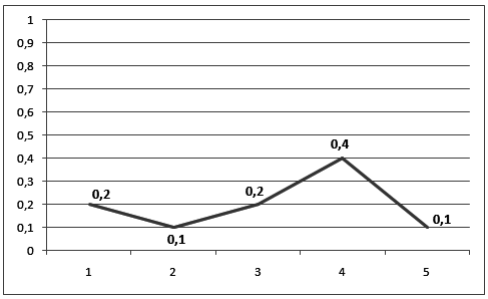
A Figura 2.8(a) ilustra o histograma de uma imagem em escala de cinza com cinco *bins* e resolução 100x100 (10000 *pixels*), enquanto que a Figura 2.8(b) apresenta o mesmo histograma, porém normalizado. Nota-se que o somatório de todos os *bins* na Figura 2.8(a) é igual a 10000, mesmo número de *pixels* na imagem original. Na Figura 2.8(b) o somatório de tais níveis é 1.0, ou seja, 100% da imagem.

Obtidos os histogramas, utilizam-se medidas de dissimilaridade (Baeza-Yates e Ribeiro-Neto, 1999) para estimar a similaridade entre dois histogramas. Duas das principais medidas de dissimilaridade são a intersecção de histogramas e a diferença absoluta de histogramas (Marques, 2011), esta última também conhecida como distância euclidiana.

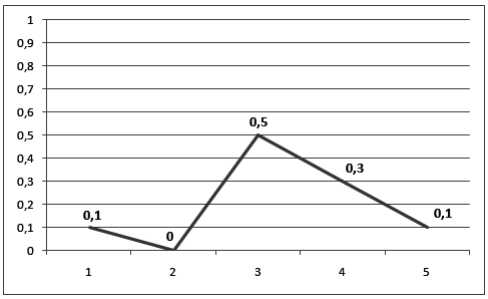
Na intersecção de histogramas, os valores de cada um dos níveis correspondentes de dois histogramas são comparados, gerando um novo histograma contendo o menor valor entre os dois níveis de histogramas analisados. Nesse caso, a intersecção é considerada como a parte comum entre dois níveis de cores, ou seja, a quantidade (proporcional ou não) de *pixels* que os dois histogramas possuem. Um exemplo gráfico da intersecção de dois histogramas normalizados é ilustrado na Figura 2.9.

Na Figura 2.9, dois histogramas de quadros diferentes (Figuras 2.9(a) e 2.9(b), respectivamente), contendo cinco níveis de cor cada, são utilizados como entrada para a intersecção de histogramas. O menor valor para cada nível é utilizado na formação de um novo histograma, apresentado na Figura 2.9(c).

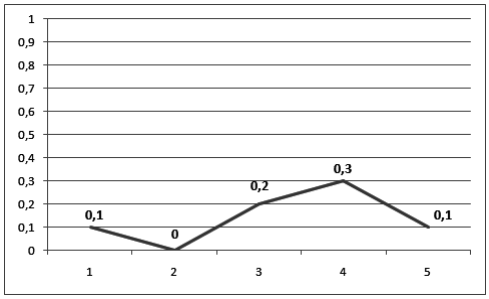
Pode-se obter, ainda, um valor único representando a intersecção entre os dois histogramas. Nesse caso, para histogramas normalizados, somam-se todos os níveis calculados através da intersecção de histogramas, resultando em um valor entre o intervalo $[0, 1]$,



(a) Histograma do quadro A.



(b) Histograma do quadro B.



(c) Histograma da intersecção dos histogramas dos quadros A e B.

Figura 2.9: Exemplo gráfico da intersecção entre dois histogramas normalizados.

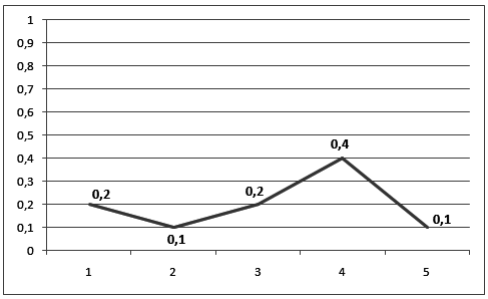
onde zero significa "nenhuma intersecção" e um significa "histogramas idênticos". Na Figura 2.9(c), o valor da intersecção resultante seria 0.7.

Já no caso da distância euclidiana, armazena-se o valor da diferença absoluta entre dois níveis correspondentes de cor de histogramas diferentes. Um exemplo gráfico da distância euclidiana entre dois histogramas é ilustrado na Figura 2.10.

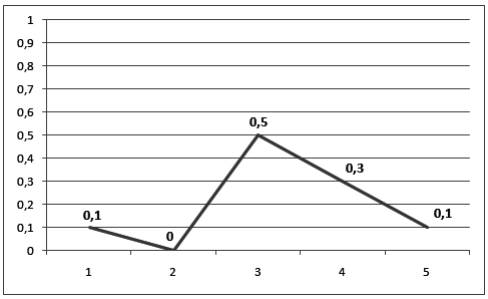
Na Figura 2.10, a distância euclidiana é calculada valor a valor nos histogramas dos quadros A e B (Figuras 2.10(a) e 2.10(b), respectivamente), resultando um em novo histograma (Figura 2.10(c)) contendo o resultado de tal operação.

Assim como no caso da intersecção de histogramas, pode-se obter um valor único da distância euclidiana entre dois histogramas. Nesse caso, para histogramas normalizados, os valores obtidos pertencem ao intervalo $[0, 2]$, onde zero significa "histogramas idênticos" e dois significa "nenhum nível de cor em comum". Na Figura 2.10(c), o valor da distância euclidiana resultante seria 0.6.

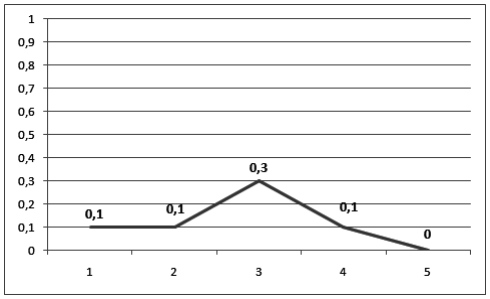
Por fim, um histograma pode ser dito global ou local. Histogramas globais utilizam como informação todos os *pixels* da imagem como base para o cálculo do histograma. Já histogramas locais utilizam grupos de *pixels*, chamados de blocos, variando em número conforme o tamanho de cada bloco. Coimbra (2011), por exemplo, utilizou em seu trabalho histogramas locais com tamanho de bloco de 16 x 16 *pixels*. Histogramas locais são mais adequados para localizar objetos em uma imagem, pois reduzem a influência do plano de



(a) Histograma do quadro A.



(b) Histograma do quadro B.



(c) Histograma resultante da distância euclidiana dos histogramas dos quadros A e B.

Figura 2.10: Exemplo gráfico da distância euclidiana entre dois histogramas normalizados.

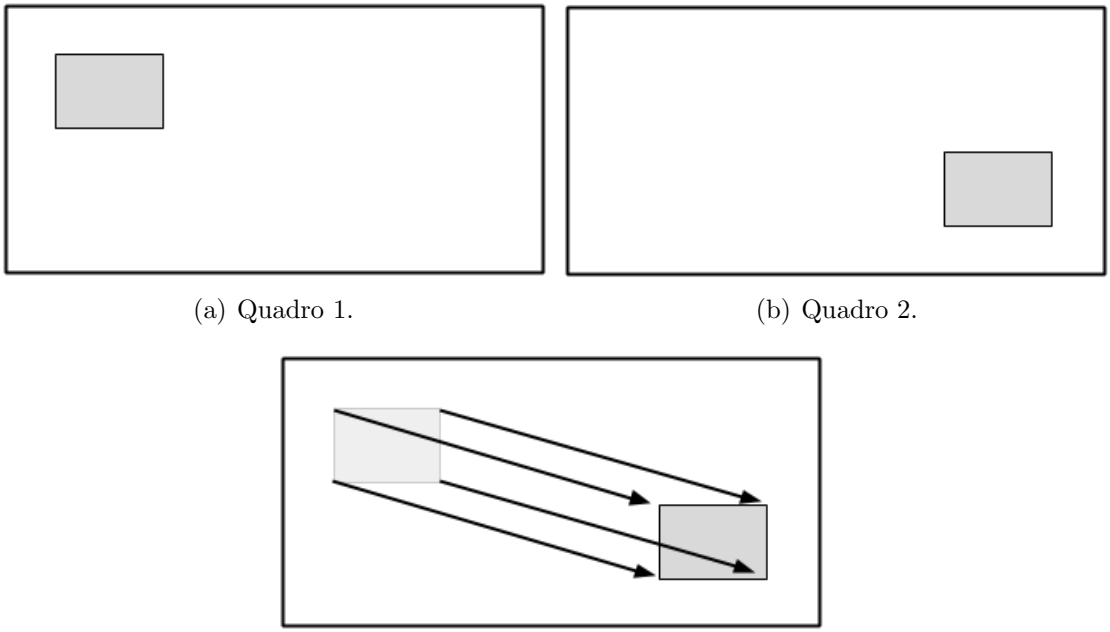
fundo (Ullman et al., 2002), enquanto que histogramas globais mostram-se mais eficientes que os histogramas locais para a segmentação em cenas (Coimbra, 2011).

2.3.2 Fluxo Óptico

Outra característica que pode ser calculada sobre um determinado vídeo é o fluxo óptico (do Inglês *optical flow*). O fluxo óptico é uma representação intermediária entre imagens que variam no tempo, onde a cada *pixel* é atribuído um vetor de velocidade que descreve o seu deslocamento temporal (Neumann, 1984). Em outras palavras, um fluxo óptico descreve o deslocamento realizado pelos *pixels* entre duas imagens que descrevem diferentes momentos temporais. Um exemplo simples de fluxo óptico é ilustrado na Figura 2.11.

O fluxo óptico pode ser obtido através de diversas técnicas diferentes, utilizando abordagens diferentes, com desempenhos distintos (Barron et al., 1992). Exemplos de técnicas diferentes incluem a técnica de Horn-Schunk (Horn e Schunck, 1980), Uras et al. (1988) e Lucas-Kanade (Lucas e Kanade, 1981).

O fluxo óptico, ao longo do tempo, foi adotado para as mais diversas áreas com objetivos distintos. Por exemplo, para encontrar humanos em vídeos de vigilância (Girisha e



(a) Quadro 1.

(b) Quadro 2.

(c) Os vetores de fluxo ópticos dos quadros 1 e 2, representando a movimentação do retângulo na imagem.

Figura 2.11: Exemplo do cálculo do fluxo óptico entre dois quadros e sua representação gráfica.

Murali, 2011), analisar as emoções humanas através da face (Duthoit et al., 2008) e até mesmo analisar e classificar o movimento fetal (Surlea et al., 2011).

Neste trabalho, utilizou-se o fluxo óptico baseado em uma implementação do algoritmo originalmente proposto por Lucas-Kanade e descrito por Bouguet (2000) para analisar a dinâmica de movimento de trechos do vídeo. Com isso, trechos de vídeos com dinâmica de movimento consideravelmente semelhantes ou, ainda, com alta ou baixa "quantidade de movimento", podem ser tratadas como semelhantes e, possivelmente, fundidas em apenas uma cena.

Maiores detalhes quanto ao cálculo do fluxo óptico, como *pixels* selecionados para análise e parâmetros utilizados para processamento podem ser encontrados na Subseção 3.1.2.

2.4 Medidas de Avaliação

A importância de medidas de avaliação reside no fato de que, com elas, torna-se possível discriminar e até mesmo comparar técnicas diferentes de segmentação de vídeo.

Diversas métricas são descritas na literatura, tais como as curvas ROC (do Inglês *Receiver Operating Characteristics*) e o Erro Quadrático Médio (do Inglês *Mean Squared*

Error - MSE) (Mood et al., 1974). Para este trabalho, adotaram-se três medidas bastante conhecidas, a precisão (*Precision*), abrangência (*Recall*) e o F-*measure* (Hua et al., 2002; Rijsbergen, 1979). Neste trabalho, utilizou-se a variação do F-*measure* conhecida como F_1, onde nem a precisão e nem a abrangência são priorizadas.

A precisão avalia a proporção de acertos na classificação sobre o total de resultados. A abrangência, por sua vez, avalia a proporção de resultados corretos obtidos em comparação com o número total de resultados corretos presentes na base e, por fim, a medida F_1 sintetiza os resultados de precisão e abrangência um único valor. Formalmente, a precisão P é definida como:

$$P = \frac{vp}{vp + fp} \tag{2.1}$$

A abrangência A, por sua vez, é definida como:

$$A = \frac{vp}{vp + fn} \tag{2.2}$$

E, finalmente, o F_1 é definido como:

$$F_1 = 2 \cdot \frac{P \cdot A}{P + A} \tag{2.3}$$

Onde vp é verdadeiro positivo, fp é falso positivo e fn é falso negativo. A descrição detalhada do significado de verdadeiro positivo, falso positivo e falso negativo no contexto de segmentação de vídeo é dada a seguir:

- **Verdadeiro positivo:** ocorre quando o método determina um momento do vídeo onde ocorre uma transição corretamente.

- **Falso positivo:** ocorre quando o método determina um momento do vídeo onde não há uma transição.

- **Falso negativo:** ocorre quando o método omite uma transição, deixando de apontá-la.

Em termos de segmentação de vídeo em tomadas ou cenas, é necessário que um especialista liste todas as transições e sua posição no fluxo de vídeo, formando a chamada base confiável (do Inglês *ground truth*). Se o algoritmo não retornar nenhum falso positivo ele é dito preciso. Caso não retorne nenhum falso negativo ele é dito abrangente.

A escolha pelas métricas de precisão, abrangência e F_1 é devido ao fato de que diversos trabalhos na área também as adotarem como medida de avaliação de seus métodos, permitindo, assim, a comparação entre a técnica desenvolvida e outras presentes na literatura. A medida F_1, em especial, foi adotada pois, além de sintetizar os valores de precisão

e abrangência em um único valor, tende a ser uma medida mais conservadora que uma média aritmética simples, por exemplo.

Como exemplo, seja um trecho de vídeo contendo dez transições de tomadas e uma transição de cena. Seja ainda uma técnica de segmentação em cenas que obtenha, para esse trecho, dez transições de cenas, uma para cada transição de tomada. Assim, essa técnica obteve 100% de abrangência, já que todas as transições de cenas foram encontradas (uma transição). Sua precisão, por outro lado, é bastante baixa: 10% (apenas um verdadeiro positivo e nove falsos positivos). Caso fosse realizada uma média simples entre a precisão e abrangência, a técnica obteria o valor de 55%, "maquiando" a baixa precisão obtida pela técnica.

Por outro lado, nas mesmas circunstâncias, a medida F_1 da técnica para a mesma segmentação seria de apenas 18.18%, um resultado bastante inferior à média simples entre a precisão e abrangência. Desta forma, caso o algoritmo apresente resultados insatisfatórios em algum quesito (precisão e/ou abrangência), o valor F_1 obtido tenderá a se aproximar do valor mais baixo, caracterizando a medida F_1 como uma média ponderada.

2.5 Técnicas de Segmentação de Vídeo

Embora a segmentação em tomadas e em cenas sejam problemas em aberto, diversas técnicas diferentes foram propostas ao longo dos anos para realizar, mesmo que de maneira limitada, tais segmentações.

Um dos problemas relacionados à segmentação de vídeo é o da seleção de quadros-chave. Normalmente, técnicas de segmentação de vídeo têm de processar um grande número de quadros de vídeo, que se traduz em métodos com alta latência. Assim, para reduzir tal volume de dados e reduzir o custo computacional de tais métodos, um subconjunto de quadros é selecionado, de maneira a representar adequadamente um segmento de vídeo. Algumas técnicas de seleção de quadros-chave e suas características são apresentadas na Subseção 2.5.1.

Quanto à segmentação em tomadas, as primeiras técnicas desenvolvidas eram centradas na detecção de transições abruptas. Atualmente, diversas técnicas no estado-da-arte atingem resultados satisfatórios em tais transições, com resultados acima de 90% em precisão e abrangência (Smeaton et al., 2010). Mais recentemente, porém, o foco dos pesquisadores voltou-se para o problema da detecção das transições graduais (Koprinska e Carrato, 2001), cuja precisão e abrangência ficam próximos a 85% em casos específicos. A Subseção 2.5.2 descreve algumas técnicas bem conhecidas e outras mais recentes para a segmentação em tomadas.

A segmentação em cenas é uma área mais recente, ainda em desenvolvimento, que busca poder representar computacionalmente a semântica latente de uma cena, um conceito subjetivo. Diversas técnicas diferentes foram propostas ao longo dos anos, com resultados diversos. A Subseção 2.5.3 apresenta alguns trabalhos relacionados de segmentação em cenas e alguns de seus resultados.

2.5.1 Algoritmos de Seleção de Quadros-Chave

Um problema relacionado às técnicas de segmentação em tomadas e em cenas é o de selecionar quadros-chave. Um quadro-chave é um quadro de um determinado segmento de vídeo, como uma tomada, por exemplo, que é utilizado como representante de tal segmento.

Os quadros-chave surgiram na área de segmentação de vídeo, e também em outras áreas, como uma forma de reduzir o número de quadros a serem analisados. Considere por exemplo que se deseje calcular a intersecção de histogramas entre duas tomadas, cada uma contendo 100 quadros. Nesse caso, seria necessário realizar 10000 (dez mil) comparações entre os histogramas. Como a operação possui ordem de crescimento quadrática, o número de comparações cresce rapidamente conforme o tamanho das tomadas, tornando o processamento inviável para um grande número de quadros de entrada.

A seleção de quadros-chaves parte do pressuposto que quadros adjacentes tendem a apresentar conteúdo visual semelhante ou até mesmo idêntico (Jacobs et al., 1995). Assim, caso sejam utilizados apenas um conjunto restrito de quadros (os "quadros-chave"), o resultado obtido pode ser idêntico ou muito próximo ao resultado que seria obtido calculando-se a intersecção entre todos os quadros das tomadas. Para obter bons resultados, é necessário que os quadros-chave sejam selecionados adequadamente. O problema da seleção de quadros-chave é exatamente como selecionar esses quadros-chave adequadamente.

Diversas técnicas de segmentação de vídeo descritas na literatura utilizam quadros-chave para representar tomadas (Chasanis et al., 2009; Rasheed e Shah, 2003, 2005; Tapu e Zaharia, 2011), principalmente técnicas de segmentação em cenas. Uma técnica bem conhecida para a seleção de quadros-chave é determinar o primeiro quadro, o último quadro ou o quadro mediano como quadro-chave (Gu et al., 2007; Li et al., 2001). Tal técnica é popular devido a sua simplicidade e ao baixo custo computacional. Nesse sentido, Souza e Goularte (2013), demonstraram que essas e outras técnicas mais rebuscadas não alcançaram resultados satisfatórios, evidenciando a necessidade de mais pesquisas para o desenvolvimento de um método que represente melhor uma tomada.

Um detalhe importante a ser considerado é de que, dependendo da tomada, apenas um quadro-chave pode não representar adequadamente a tomada. Em algumas tomadas, por

exemplo, o cenário pode apresentar mudanças significativas conforme o tempo, tornando a seleção de apenas um quadro-chave limitada. Assim, costuma-se selecionar um conjunto de quadros-chave para cada tomada (Besiris et al., 2007; Liu et al., 2004; Mukherjee et al., 2007; Rasheed e Shah, 2003, 2005; Rathod e Nikam, 2013). O desafio para essas técnicas, além de selecionar quadros adequados, é o de reduzir a redundância entre o conjunto de quadros-chave selecionados.

Alguns trabalhos presentes na literatura procuram determinar as melhores técnicas de seleção de quadros-chave (Pickering e Rüger, 2003; Souza e Goularte, 2013; Sujatha e Mudenagudi, 2011). Porém, tais análises são geralmente limitadas, já que se utiliza de métricas diferentes e comparam um número muito restrito de técnicas. Assim, devido à ausência de uma técnica consensualmente considerada como "adequada", é usual que técnicas de segmentação desenvolvam métodos próprios de seleção de quadros-chaves (Besiris et al., 2007; Chasanis et al., 2008; Pimentel Filho et al., 2008).

Assim, neste trabalho, desenvolveu-se um método de seleção de quadros-chave que procura encontrar um conjunto de quadros-chaves diferentes entre si e complementares, de maneira a representar adequadamente tomadas complexas. O método de seleção de quadros-chave é descrito na Subseção 3.2.1.

2.5.2 Algoritmos de Segmentação em Tomadas

Os algoritmos de segmentação, em geral, podem ser classificados quanto ao tipo de entrada esperada: um vídeo comprimido ou descomprimido. Técnicas de segmentação de vídeos comprimidos utiliza-se de informações como coeficientes DCT (*Discrete Cosine Transform*) (Ahmed et al., 1974), vetores de movimento, entre outros, para auxiliar a segmentação de vídeo (Rasheed e Shah, 2003; Xiang-wei et al., 2009).

Usualmente, as técnicas de segmentação em tomadas trabalham com vídeos descomprimidos através da comparação quadro-a-quadro (Koprinska e Carrato, 2001). Quando dois quadros consecutivos são suficientemente dissimilares, uma transição abrupta é detectada. Para as transições graduais, utilizam-se técnicas mais sofisticadas que medem a dissimilaridade entre um conjunto de quadros. Nesse sentido, o foco deste trabalho são técnicas de segmentação de vídeos descomprimidos, dado que a técnica desenvolvida também atua sobre vídeos descomprimidos.

Tradicionalmente, as técnicas de segmentação em tomadas basearam-se em análises tais como de *pixels*, blocos de *pixels* e histogramas. Porém, diversas técnicas foram propostas utilizando-se de *clustering* (Ferman e Tekalp, 1998; Günsel et al., 1998), modelos (Aigrain e Joly, 1994; Hampapur et al., 1994) e também utilizando outras características do vídeo, como a análise da intensidade das bordas das imagens (Zabih et al., 1999).

Na técnica de segmentação em tomadas pela análise de *pixels*, analisam-se as diferenças entre *pixels* correspondentes em quadros adjacentes. Uma das formas de análise é calcular a soma absoluta das diferenças entre os *pixels* (Kikukawa e Kawafuchi, 1992). Nesse caso, se o valor encontrado for maior que um limiar especificado anteriormente, uma transição abrupta é detectada. Porém, como desvantagem, mudanças significativas em partes do quadro podem causar o extrapolamento do limiar, tornando-a sensível aos movimentos da câmera e dos objetos. Algumas melhorias foram propostas, como a introdução de outro limiar para contabilizar a quantia de *pixels* que apresentaram mudanças significativas na intensidade ou cor (Zhang et al., 1993). Porém, mesmo com tais melhorias, a técnica ainda é sensível à movimentação da câmera e dos objetos.

A técnica de comparação de blocos de *pixels* surgiu como uma forma mais robusta que a comparação de *pixels* em relação à movimentação dos objetos e da câmera. Nessa técnica, um quadro é dividido em blocos de tamanho definido e comparado com o bloco correspondente no quadro adjacente. Diversas formas de comparação de blocos foram propostas (Lefèvre et al., 2003; Shahraray, 1995), porém, a técnica ainda é sensível à movimentação da câmera e/ou dos objetos, além de ter um custo de processamento elevado em algumas formas de calcular as diferenças entre blocos, como o apresentado por Katsuri e Jain (1991).

Já a segmentação baseada em histogramas apresenta como ideia principal que, dado dois quadros adjacentes com fundo igual e os mesmos objetos (em movimento ou não), os histogramas dos dois quadros serão semelhantes (Koprinska e Carrato, 2001). Como vantagem, cita-se que histogramas são invariantes à rotação, escala e diferenças de ângulos (Swain, 1993). Os histogramas podem tanto ser locais, extraídos de uma parte do quadro, como globais, representando o quadro por inteiro. Diversas técnicas foram propostas tanto para histogramas locais (Fu e Zeng, 2009; Lee e Man-Ching, 1994; Rathod e Nikam, 2013; Swanberg et al., 1993) como para globais (Gargi et al., 1995; Sowjanya e Mishra, 2012; Zhang et al., 1993). Como desvantagens, cita-se que duas imagens com conteúdo totalmente diverso podem ter histogramas iguais, porém, tanto a chance desse caso ocorrer é baixa como técnicas para tratar tal problema já foram apresentadas (Pass e Zabih, 1999). Além disso, técnicas baseadas em comparação de histogramas apresentam erros em casos de mudanças bruscas de tonalidade, como *flashes* (Jiang et al., 1998). Por fim, cita-se que histogramas são sensíveis a grandes movimentos de câmera (Geetha e Narayanan, 2008).

Mesmo com tais problemas, Geetha e Narayanan (2008), destacam que a diferença de histogramas era a técnica mais utilizada para o cálculo de dissimilaridade entre dois quadros adjacentes. Infelizmente, não foram encontrados trabalhos mais recentes que procuram identificar a(s) técnica(s) mais utilizada(s) para a segmentação em tomadas. Até aquele momento, o histograma era amplamente utilizado, pois, mesmo com o desen-

volvimento de outras técnicas invariantes a problemas de iluminação, tais técnicas não conseguem obter uma relação custo/benefício significativamente melhor que as técnicas de comparação de histogramas (Yuan et al., 2007).

Alguns trabalhos desenvolvidos têm como objetivo tratar algumas das falhas inerentes à técnica adotada, como por exemplo, o trabalho de Hong-Cai et al. (2010), onde os autores procuram reduzir o impacto que *"flashes"* e a presença de textos têm sobre a segmentação baseada em histogramas, através do uso de limiares globais e adaptativos, além da detecção de tipos de *flash*.

Há também trabalhos que incluem o uso de técnicas diferentes para a detecção de transições abruptas e graduais. Yu, Tian et al. (2007), especificam o uso de uma medida chamada de *Shot Boundary Coefficient*, obtida através da diferença média de janelas deslizantes, combinado com histogramas do espaço de cor HSV. Para detectar mudanças abruptas de tomadas é utilizada uma rede *Self-Organizing Map* (Kohonen, 1990) para a transição gradual. Segundo os autores, a técnica apresenta uma boa relação entre velocidade e precisão, além de certa tolerância a ruído.

Outras técnicas baseiam-se em parâmetros pré-fixados para a análise e segmentação do vídeo. Tais algoritmos são chamados "paramétricos", necessitando de conhecimento prévio de detalhes únicos do vídeo, do domínio, do formato, entre outros. Nesse sentido, Manjunath et al. (2011) propõe uma técnica não paramétrica, independente de conhecimento prévio do vídeo, em que os parâmetros são calculados durante a execução do processo segmentador. Outras abordagens semelhantes podem ser encontradas nos trabalhos de Koumaras et al. (2006) e Punitha e Jose (2009).

Outras técnicas de segmentação em tomadas, em estado-da-arte, foram submetidas ao TRECVid. O evento disponibilizou, de 2001 a 2007 (inclusive), uma grande base de vídeos de teste, uma métrica uniformizada de comparação e um fórum para as diferentes organizações compararem seus resultados (Smeaton et al., 2006). Nesses anos, o TRECVid disponibilizou uma trilha de avaliação de técnicas de segmentação de vídeo em tomadas onde diferentes algoritmos foram aceitos para a avaliação de, na média, seis horas de vídeos pré-definidos em busca dos valores de precisão, abrangência e F-*measure*.

Os resultados sintetizados do TRECVid de 2005 (Smeaton et al., 2010), por exemplo, mostram que as melhores técnicas atingiram valores de F-*measure* de cerca de 90% para transições abruptas. Para transições graduais, porém, o índice de F-*measure* apresentou valores de 68% a 79%. Nota-se que a maioria das técnicas de segmentação fizeram uso de histogramas de cor (15 técnicas de 21), aprendizagem de máquina (10 técnicas de 21) e detecção de *flashes* (11 técnicas de 21), sendo que diversas técnicas utilizaram mais de uma abordagem ao mesmo tempo. É interessante notar, ainda, que apenas uma das 10 melhores técnicas para a segmentação em transições abruptas não utilizou aprendizagem

de máquina. Para transições graduais, três técnicas não utilizaram aprendizagem de máquina para realizar a segmentação em tomadas, indicando o predomínio de tais técnicas entre as que obtêm os melhores índices de precisão e abrangência para a segmentação em tomadas.

Porém, como desvantagem, cita-se que algoritmos de aprendizagem de máquina usualmente necessitam de treinamento prévio com bases semelhantes aos vídeos que serão segmentados, processo não trivial, principalmente para vídeos complexos com diversos contextos ou características presentes.

2.5.3 Algoritmos de Segmentação em Cenas

A segmentação de vídeo em cenas de alto nível semântico é tanto a que provê maior informação a aplicações voltadas para usuário, devido à similaridade do conceito subjetivo de "cena" com os segmentos retornados por técnicas para a segmentação em cenas, como também a que apresenta maiores dificuldades de ser alcançada. A dificuldade primordial da segmentação em cenas é a de que as cenas são limitadas semanticamente, enquanto que tomadas são limitadas fisicamente (Sakarya e Telatar, 2010).

As diversas técnicas de segmentação em cenas podem ser divididas em três abordagens (Gu et al., 2007; Wang et al., 2008b): baseadas em união, baseadas em divisão e baseadas em modelos estatísticos. As três classificações são apresentadas a seguir:

- **Baseadas em união (*merging*):** são consideradas as técnicas tradicionais na segmentação em cenas. Algoritmos dessa abordagem primeiro dividem o vídeo em tomadas que são progressivamente unidas formando as cenas. É conhecida também como técnica *bottom-up*. Exemplos de tal abordagem são encontrados em trabalhos como Chu et al. (2009) e Papadopoulos et al. (2008).

- **Baseadas em divisão (*splitting*):** método onde as cenas são criadas diretamente através da divisão do vídeo em estilo *top-down*. Um exemplo de tal abordagem é o trabalho apresentado por Rasheed e Shah (2005), que constroem um grafo de similaridade do vídeo e o particiona usando cortes normalizados. Os sub-grafos resultantes são as cenas do vídeo. Tavanapong e Zhou (2004), por sua vez, apresentam uma definição para filmes narrativos e realizam a clusterização de tomadas seguindo tal definição.

- **Baseadas em modelos:** nessa abordagem, são adotados modelos estatísticos para determinar a similaridade de cenas. O trabalho apresentado por Zhai e Shah (2005b), por exemplo, baseia-se no uso de um modelo estocástico de Monte Carlo para simular a geração de cenas para determinar as bordas das cenas. Uma desvantagem importante para tal tipo de técnica é a limitação ao modelo adotado e a

necessidade de o vídeo seguir o modelo, assim, vídeo de domínios como esportes e telejornalismo podem ser detectados com bom desempenho, mas vídeos do domínio de filmes e vídeos caseiros apresentam um baixo desempenho (Gu et al., 2007).

Tal classificação de segmentação em cenas, porém, não é consensual. Hu et al. (2011), por exemplo, citam uma quarta abordagem para a segmentação em cenas, baseada na classificação das bordas da tomada. Em tal abordagem os limites de cada tomada são classificados, de acordo com o cálculo de algumas características, em "limite" de cena ou "não limite" de cena. Goela et al. (2007) apresentam, por exemplo, um método de detecção de cenas em vídeos transmitidos em *broadcast*, tais como sinais de TV. Na abordagem, *Support Vector Machine* (SVM) é usado para classificar os limites das tomadas, usando limiares de cenas pré-segmentadas manualmente como exemplos positivos e negativos para o SVM.

Uma técnica bem difundida de segmentação em cenas é baseada em medidas de comparação tomada-a-tomada. Kender e Yeo (1998) nomeou tais técnicas como baseadas em "coerência entre tomadas". Em tais trabalhos, uma medida de dissimilaridade é estimada entre duas tomadas quaisquer, chamada de coerência, e, após uma análise do valor encontrado, as tomadas são agrupadas em cenas. Embora inicialmente fosse utilizada apenas a análise da coerência entre tomadas baseada em histogramas (Kender e Yeo, 1998), abordagens posteriores (Bai et al., 2008; Rasheed e Shah, 2003) adotam, também, métricas capazes de identificar a dinâmica de movimento para a fusão de tomadas em cenas. Uma das vantagens de tais técnicas é o baixo custo computacional, o fato de não necessitar de conhecimento prévio do domínio do vídeo de entrada e nem de treinamento prévio (como em técnicas de aprendizagem de máquina) e, ainda assim, apresentar resultados significativos (Rasheed e Shah, 2003).

Uma abordagem relativamente nova de segmentação em cenas é baseada em multimodalidade. O "modo" refere-se ao tipo de mídia de entrada, assim, técnicas multimodais se aproveitam de diversos tipos de mídia para realizar a segmentação. O objetivo de se utilizar múltiplas fontes de informação, como o fluxo de vídeo e de áudio, por exemplo (Coimbra e Goularte, 2009), é o de um determinado tipo de dado suprir as deficiências de outro tipo de dado. Um exemplo de uma técnica multimodal de segmentação em cenas é o trabalho apresentado por Poulise e Moens (2011), que criaram um segmentador de cenas das Olimpíadas de 2008 usando *closed captions* e o próprio fluxo de vídeo. O método utiliza tanto uma técnica desenvolvida para segmentação de textos aplicada em um trabalho anterior (Poulisse e Moens, 2011) como a técnica para a obtenção de informações visuais através de um núcleo espaço-piramidal (do Inglês *spatial-pyramid kernel*) (Lazebnik et al., 2006). Segundo os autores, a justificativa de utilizar o domínio de vídeos olímpicos é de que são diferentes dos vídeos esportivos utilizados em outros trabalhos (Ariki et al.,

2003; Xie et al., 2004), pois possuem cenas de preparação, concentração, entre outros. Diferenciando-se, também, de vídeos jornalísticos, foco de outras técnicas (Hsu e Chang, 2004; Zhai e Shah, 2005a), já que não existem as regras de produção e o *closed caption* tem menor poder de representação e expressividade.

Em outro trabalho que adota a multimodalidade, Sundaram e Chang (2000) detectaram cenas utilizando-se, separadamente, dos fluxos de vídeo e de áudio, unindo-as através de um algoritmo de vizinho mais próximo (do Inglês *nearest neighbor*) com algumas restrições temporais. Tal abordagem é adotada por outros trabalhos, tal como o de Pfeiffer et al. (2001). Como desvantagem, Hu et al. (2011) destacam a dificuldade de correlacionar informações de segmentos de áudio com tomadas visuais.

Na literatura, diversos trabalhos mostram a grande variedade de abordagens adotadas com o intuito de melhorar o desempenho da segmentação em cenas. O trabalho apresentando por Sakarya e Telatar (2010), por exemplo, adota um sistema de clusterização não supervisionado tendo por base informações filtradas da matriz de similaridades obtidas na comparação tomada-a-tomada. Outras abordagens semelhantes podem ser encontradas em (Rasheed e Shah, 2005; Sakarya e Telatar, 2007).

Yun e Shah (2006) apresentam uma técnica onde demarcadores de limites de cenas são dispostas randomicamente pelo vídeo. Através de análises estatísticas, como proximidade dos dados, são determinados os valores finais dos limiares, resultando na segmentação do vídeo em cenas.

Hanjalic (2002) propõe uma técnica de detecção de limiares da cena, chamado por ele de *story units*, em filmes baseadas em uma medida de similaridade de blocos entre quadros-chave de cada tomada.

Chasanis et al. (2009) desenvolveram uma técnica de detecção de tomadas baseadas em histogramas de cor, seguida por uma extração de quadros-chave com base na clusterização espectral, obtendo-se a segmentação em cenas, finalmente, pela clusterização das tomadas baseadas na similaridade de cores. Segundo os autores, a técnica proposta obteve bons resultados em comparação com outros trabalhos semelhantes tais como os apresentados por Rasheed e Shah (2005) e Yeung et al. (1998).

Sakarya e Telatar (2008) apresentaram um método onde tomadas são extraídas do vídeo e representadas como vértices de um grafo. As arestas entre os vértices são calculados através de uma medida de similaridade espaço-temporal. A partir daí, é detectada uma única cena que apresenta o maior grau de similaridade entre tomadas e o maior grau de dissimilaridade com as demais tomadas. As demais cenas são encontradas usando percursos em árvore. Segundo os autores, o resultado é superior ao trabalho publicado por Rasheed e Shah (2005), principalmente devido à menor taxa de falsos negativos obtida.

Sawai *et at.* (2011) utilizam uma série de cenas segmentadas manualmente, aliadas a características extraídas do vídeo de entrada como detecção de faces (Viola e Jones, 2001), detecção de *flashes* (Takimoto et al., 2006) e a intensidade dos *pixels* para detecção de cenas específicas de uma festa de casamento. Segundo os autores, a taxa de acerto do algoritmo é de cerca de 83%, alcançando cerca de 96% em alguns casos.

Tapu et al. (2011) propõem um *framework* para segmentação em tomadas e em cenas utilizando-se de diversas abordagens. Primeiro, uma comparação entre histogramas HSV de quadros sucessivos é realizado para determinar transições abruptas. As transições graduais e quadros que apresentaram resultados não conclusivos com a comparação de histogramas são submetidos a um método de segmentação baseado no modelo de particionamento de grafos (Yuan et al., 2007). Após a seleção de quadros-chave para cada tomada, o algoritmo procura dividir as tomadas em *clusters* que formarão as cenas. Segundo os autores, a abordagem consegue atingir 68% de precisão e 87% de abrangência no domínio de filmes hollywoodianos.

Embora interessantes, as diversas técnicas citadas apresentam limitações importantes, incluindo:

- Alto custo computacional, como nos trabalhos de (Chasanis et al., 2009; Sakarya e Telatar, 2010; Tapu e Zaharia, 2011), graças ao uso de algoritmos de clusterização e em (Zhai e Shah, 2006) devido ao cálculo simultâneo de diversas cadeias de Markov e o uso de análises estatísticas.

- Limitação a determinados domínios, como nos trabalhos publicados por Poulisse e Moens (2011) e Sawai et al. (2011), restritos aos domínios "jogos olímpicos" e "festas de casamento", respectivamente.

- Necessidade de intervenção de especialista, como no trabalho apresentado por Sawai et al. (2011).

- Baixa confiabilidade, como no trabalho publicado por Hanjalic (2002), onde foi utilizada a técnica de similaridade entre blocos, conhecida tanto pelo alto custo computacional como pelo fato de blocos diferentes poderem retornar valores iguais.

- Necessidade de outros tipos de informações além do vídeo, como fluxo de áudio (Pfeiffer et al., 2001; Sundaram e Chang, 2000) e de *closed captions* (Poulisse e Moens, 2011), que podem não estar disponíveis.

- Limitações quanto ao tipo de fluxo de vídeo de entrada (Rasheed e Shah, 2003).

Assim, embora uma ampla gama de trabalhos na literatura trate da segmentação em cenas, o problema de segmentar um vídeo em cenas continua sem uma solução definitiva,

sendo que as soluções apresentadas, ou apresentam um desempenho insatisfatório, ou são limitados pelo domínio, necessitando parâmetros ou informações além do fluxo de vídeo. Nesse sentido, a técnica baseada em coerência de tomadas, ou simplesmente CV, apresenta-se como uma alternativa importante às demais técnicas devido ao baixo custo computacional.

Infelizmente, mesmo a técnica em estado-da-arte que faz uso da CV, a técnica desenvolvida por Rasheed e Shah (2003), possui como limitação o formato de entrada do vídeo, como citado anteriormente. Nota-se que a técnica desenvolvida em 2003 ainda ser considerada como estado-da-arte evidencia que as demais técnicas de segmentação em cenas, como as baseadas em aprendizagem de máquina e/ou multimodalidade, receberam maior atenção dos pesquisadores.

A falta de maiores estudos sobre a CV deve-se, provavelmente, ao fato de que técnicas baseadas em aprendizagem de máquina ou multimodalidade apresentaram resultados mais significativos, fazendo os pesquisadores da área concentrar-se em estudar tais técnicas. Porém, a CV tem como vantagens o baixo custo computacional e a possibilidade de ser utilizada em outros tipos de técnicas (como as baseadas em aprendizagem de máquina e multimodalidade), o que a torna um importante foco de pesquisa para a área de segmentação de vídeo em cenas.

2.6 Considerações finais

A expansão dos meios de acesso e produção de conteúdo multimídia, principalmente vídeo, fez com que a área de segmentação de vídeo ganhasse importância, devido à necessidade de localizar, agrupar, armazenar e selecionar um volume cada vez maior de dados. O advento do vídeo digital, criado na segunda metade do século XX, trouxe consigo uma ampla gama de características de vídeo, algumas das quais descritas neste capítulo. Infelizmente, embora em grande número, tais características representam apenas informações de baixo nível semântico, que possuem pouco significado para a maioria dos usuários. Assim, a segmentação de vídeo digital procura, utilizando as características de baixo nível semântico que são extraídas ou calculadas do vídeo, retornar ao usuário um trecho ou segmento de vídeo com maior valor semântico, ou seja, que possua algum significado discernível pelo usuário.

Este capítulo apresentou, nesse sentido, dois métodos principais para a segmentação de vídeo: a segmentação em tomadas e a segmentação em cenas. A segmentação em tomadas, explorada intensamente nas décadas de 1990 e 2000, obtém resultados bastante satisfatórios para diversos tipos de vídeos, com precisão acima de 80% mesmo no pior caso. Já a segmentação em cenas, uma área de pesquisa mais recente, ainda carece de

algoritmos capazes de entregar bons resultados no caso geral (por exemplo, em vídeos de qualquer domínio ou formato).

Das técnicas de segmentação em cenas existentes, duas delas destacam-se: as baseadas em aprendizagem de máquina e as baseadas em multimodalidade. Embora apresentem resultados superiores aos resultados obtidos pela técnica baseada em CV, ambas sofrem de problemas importantes: a necessidade de treinamento prévio com vídeos similares ao vídeo que se deseja segmentar e a necessidade de obter informações de mídias diferentes, respectivamente. Além disso, tais técnicas possuem alto grau de complexidade, ocasionando um elevado custo computacional, devido principalmente à fase de treinamento (aprendizagem de máquina) e o de processar múltiplas mídias (multimodalidade).

Nesse contexto, a técnica de segmentação em cenas baseadas em características visuais apresenta duas vantagens importantes: não necessita de treinamento prévio e de outras mídias além do fluxo de vídeo. Graças a sua relativa simplicidade, o custo computacional de tal técnica é consideravelmente menor que as técnicas baseadas em aprendizagem de máquina ou baseadas em multimodalidade. Por fim, uma nova técnica de segmentação de vídeo em cenas baseada em CV poderia ser adotada por outras técnicas, principalmente devido ao baixo custo computacional e a independência do domínio ou formato de entrada do vídeo.

A técnica de CV, embora não tenha sido o foco de pesquisas na última década, ainda pode ser melhorada com a introdução de melhores métodos de cálculo da CV, além do desenvolvimento de métodos capazes de reduzir os problemas enfrentados pela CV, como por exemplo, a adoção de uma análise de movimento entre segmentos de vídeo. Nesse sentido, uma técnica baseada na CV para a segmentação de vídeo digital em cenas é apresentada no Capítulo 3.

Capítulo

3

A Técnica de Segmentação

A técnica de segmentação proposta é baseada no conceito de coerência visual (apresentada na Subseção 2.5.3), onde é estimada uma medida de "coerência" de um dado segmento em relação a um conjunto definido de segmentos anteriores. A técnica utiliza duas características básicas, os histogramas e o fluxo óptico, descritos na Seção 2.3, com o intuito de oferecer um bom desempenho e independência ao formato de entrada do vídeo.

Com o objetivo de reduzir a complexidade de implementação, diminuir o acoplamento e promover a modularidade da implementação, a técnica foi desenvolvida em duas fases: a segmentação em tomadas e a segmentação em cenas. A Figura 3.1 ilustra a técnica dividida em duas fases e os procedimentos realizados em cada uma das fases.

Como apresentado na Figura 3.1, na primeira fase são extraídas as características do vídeo e é realizada a segmentação em tomadas. Na segunda fase, as características extraídas na primeira fase do algoritmo são utilizadas para selecionar os quadros-chave e determinar a coerência entre as tomadas para, assim, realizar progressivamente a fusão entre as tomadas, formando as cenas.

A primeira fase da técnica, a segmentação em tomadas, é descrita na Seção 3.1. Já a segunda fase da técnica, a segmentação em cenas, é descrita na Seção 3.2. Adicionalmente, uma implementação da técnica é descrita na Seção 3.3, utilizada na avaliação da técnica apresentada no Capítulo 4. Por fim, na Seção 3.4, discute-se uma série de considerações finais quanto à técnica desenvolvida.

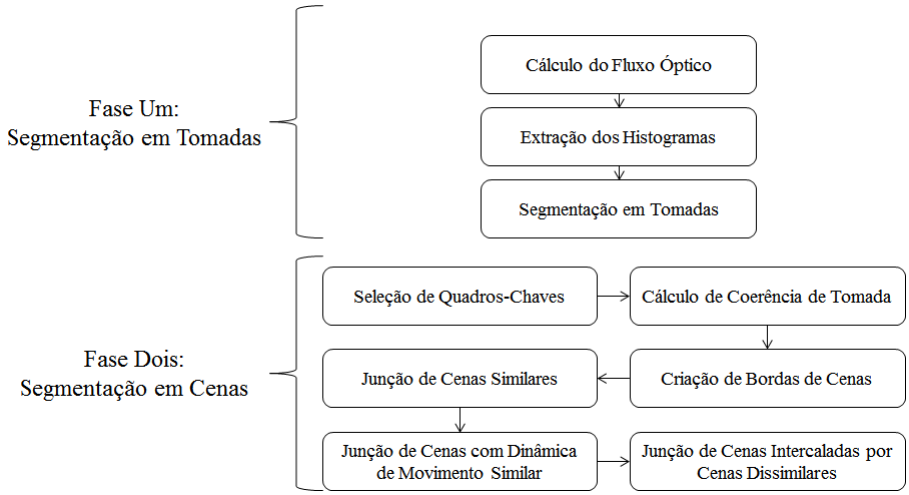


Figura 3.1: Divisão da técnica de segmentação em duas fases: a segmentação em tomadas e a segmentação em cenas.

3.1 Fase Um: Segmentação em Tomadas

A primeira fase do algoritmo inclui os procedimentos necessários para a detecção das transições de tomadas, aliado ao cálculo do fluxo óptico dos quadros do vídeo.

Nessa fase, o vídeo é decodificado, se necessário, para obter os quadros do vídeo. Com os quadros, é possível realizar tanto a extração dos histogramas como o cálculo do fluxo óptico. Com os histogramas extraídos, pode-se proceder com a segmentação em tomadas propriamente dita.

A extração dos histogramas é detalhada na Seção 3.1.1, o cálculo do fluxo óptico é descrito na Seção 3.1.2 e, por fim, o procedimento de segmentação em tomadas é apresentado na Subseção 3.1.3.

3.1.1 Extração dos Histogramas

Neste trabalho os histogramas são gerados tendo como base quadros no espaço de cor HSV (Smith, 1978). Tal espaço de cor foi adotado, pois se trata de um modelo baseado no sistema visual humano, apresentar resultados superiores de precisão e abrangência em relação ao espaço de cor RGB para segmentação de vídeo (Jeong, 2001) e também ser utilizado em diversos trabalhos da área (Bai et al., 2008; Rasheed e Shah, 2005; Yu et al., 2007b).

Assim como em trabalhos relacionados (Rasheed e Shah, 2003), foi utilizada a quantização dos histogramas HSV com 16 *bins* na proporção: 8 para *hue*, 4 para *saturation* e 4 para *value*, caracterizando uma subamostragem de crominância (Gonzalez e Woods, 2007). No total, o histograma calculado possui 128 diferentes níveis de cor.

O fluxograma, simplificado, do procedimento de extração dos histogramas é ilustrado na Figura 3.2.

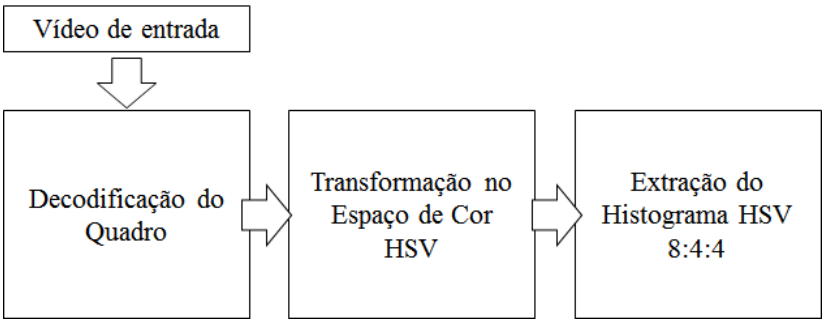


Figura 3.2: Fluxograma simplificado descrevendo os procedimentos realizados para a extração dos histogramas dos quadros do vídeo.

Na Figura 3.2, a decodificação é um processo já estabelecido para os mais diversos formatos de vídeo de entrada, com bibliotecas disponíveis nos mais diversos sistemas, como a FFMpeg[1]. Nota-se que a decodificação só é realizada com o objetivo de se obter os quadros do vídeo. Caso o vídeo de entrada não seja compactado, o processo de decodificação apenas separa o fluxo de entrada nos quadros individuais.

Ao final do processo de decodificação, tem-se como resultado todos os quadros do vídeo no espaço de cor RGB, na resolução especificada no *stream* do vídeo de entrada. O processo de conversão do espaço de cor RGB para HSV, por sua vez, é realizado utilizando-se o procedimento definido pela biblioteca OpenCV[2], descritos nas Equações 3.1, 3.2 e 3.3.

$$V \leftarrow \max(R, G, B) \tag{3.1}$$

Primeiro, calcula-se o valor de *Value* (V), utilizado como base tanto no cálculo do valor de *Saturation* (S) e de *Hue* (H). O valor S é igual ao maior valor entre os canais de cor *red* (R), *green* (G) e *blue* (B).

$$S \leftarrow \frac{V - \min(R, G, B)}{V} \tag{3.2}$$

O segundo cálculo refere-se ao valor de *Saturation* (S). Caso o valor V calculado na Equação 3.1 seja zero (ou seja, o valor RGB seja 0 em todos os canais, ou um *pixel* na cor preta), considera-se que o valor de S também será zero.

[1]http://www.ffmpeg.org/
[2]http://opencv.org/

$$H \leftarrow \begin{cases} \frac{60(G-B)}{V-min(R,G,B)} & \text{caso } V = R \\ \frac{120+60(B-R)}{V-min(R,G,B)} & \text{caso } V = G \\ \frac{240+60(R-G)}{V-min(R,G,B)} & \text{caso } V = B \end{cases} \tag{3.3}$$

O terceiro cálculo refere-se ao valor de *Hue* (H). Nesse caso, o cálculo difere-se de acordo com o canal de cor escolhido no cálculo do valor V, apresentado na Equação 3.2.

O processo de conversão RGB->HSV tem por objetivo separar as informações de cor (crominância) da informação de iluminação (luminância). Assim, ao adotar a quantização 8:4:4, prioriza-se a luminância, que é mais bem percebida pelo sistema visual humano (Gonzalez e Woods, 2007), que pode possivelmente aumentando a qualidade subjetiva dos resultados obtidos.

Por fim, o histograma de cada quadro é determinado através da contagem dos *pixels* pertencentes a cada *bin* do histograma HSV.

3.1.2 Cálculo do Fluxo Óptico

Na técnica de segmentação descrita nesse trabalho, o cálculo do fluxo óptico, apresentado na Seção 2.3.2, têm por objetivo quantificar o movimento em trechos específicos de vídeo. Para isso, o fluxo óptico é calculado a cada dois quadros decodificados, obtendo-se um conjunto de vetores que representam a movimentação presentes entre aqueles quadros.

O fluxograma, simplificado, do procedimento de cálculo do fluxo óptico é ilustrado na Figura 3.3.

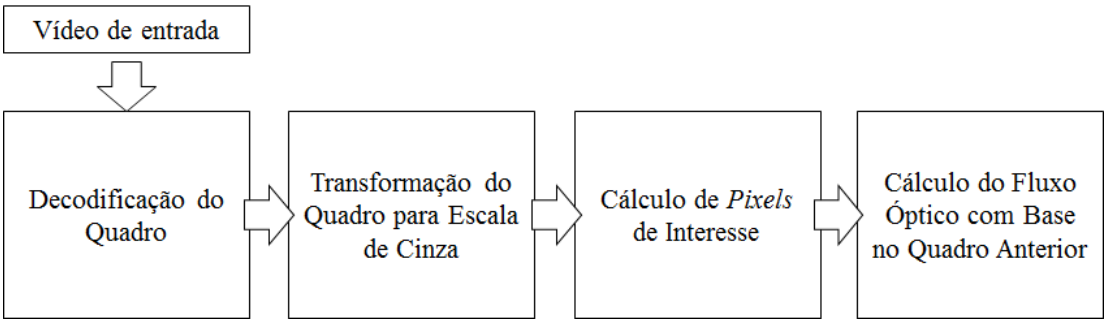


Figura 3.3: Fluxograma simplificado descrevendo os procedimentos realizados para a criação dos vetores de fluxo óptico dos quadros do vídeo.

Para o cálculo, consideram-se apenas quadros adjacentes. Assim, o primeiro quadro do vídeo não possui qualquer vetor de fluxo óptico. A partir do segundo quadro do vídeo, o fluxo óptico é calculado considerando-se o quadro "atual" e o quadro imediatamente anterior, que deve ser armazenado temporariamente a cada quadro decodificado.

Como foi utilizada uma implementação do algoritmo de Lucas-Kanade (Bouguet, 2000) para o cálculo do fluxo óptico, diversos parâmetros tiveram que ser estimados empiricamente. Os principais parâmetros utilizados para o cálculo do fluxo óptico e seus valores são apresentados a seguir:

- Tamanho de janela de busca para *pixels* similares: 15x15.
- Número máximo de *pixels* significativos a serem procurados: 1000.
- Máxima diferença entre dois conjuntos de *pixels* : 3%.
- Número máximo de iterações do algoritmo: 20 iterações.

Assim, dado um *pixel* qualquer no quadro Q_x, procura-se o mesmo *pixel* no quadro Q_{x-1} em uma janela de busca de 15x15 a partir do ponto de origem, buscando até 1000 *pixels*, parando a comparação de um *pixel* individual quando o número de iterações alcançar 20 ou a diferença mínima entre os *pixels* correspondentes for 3% (o que ocorrer primeiro).

Para acelerar o processo de cálculo do fluxo óptico adota-se um conjunto determinado de *pixels* para serem analisados. Tal abordagem parte do pressuposto que a movimentação pode ser detectada e quantificada através da movimentação de parte dos *pixels*, não sendo necessário analisar todos os *pixels* de um quadro.

Na técnica descrita, adotam-se alguns dos *pixels* formadores das bordas da imagem como *pixels* aptos a passarem pelo procedimento de cálculo do fluxo óptico. Para isso, são selecionados os *pixels* de borda com elevado autovalor (Chen et al., 2010) em vizinhanças de 3x3. Com isso, são detectadas as bordas significativas dos objetos presentes nos quadros, eliminando-se a necessidade de processar todos os *pixels* do quadro.

Deve-se destacar que o quadro a ser processado pelo procedimento de extração do fluxo óptico é convertido para a escala-de-cinza, haja vista que as bordas dos objetos são mantidas e pequenas variações de iluminação ou cor são eliminadas, ocasionando o aumento da velocidade de processamento.

O resultado do processamento é armazenado na forma de tuplas, onde o primeiro elemento é a posição no eixo X,Y de determinado *pixel* no quadro anterior e o segundo elemento é a posição no eixo X,Y do mesmo *pixel* no quadro "atual".

Por fim, destaca-se que o fluxo óptico, embora seja calculado na primeira fase do algoritmo, não será utilizado até os procedimentos finais da segunda fase do algoritmo. Assim, o fluxo óptico é calculado junto a primeira fase para evitar o processo de abertura e decodificação do vídeo de entrada na segunda fase da técnica. Com isso, todas as características do vídeo são calculadas nessa fase, única a ter acesso direto ao vídeo de entrada.

3.1.3 Segmentação em Tomadas

Com os histogramas de todos os quadros extraídos, a técnica pode realizar a segmentação em tomadas baseada na comparação quadro-a-quadro dos histogramas HSV. A Figura 3.4 ilustra o fluxograma, simplificado, dos procedimentos realizados com o intuito de se obter a segmentação em tomadas.

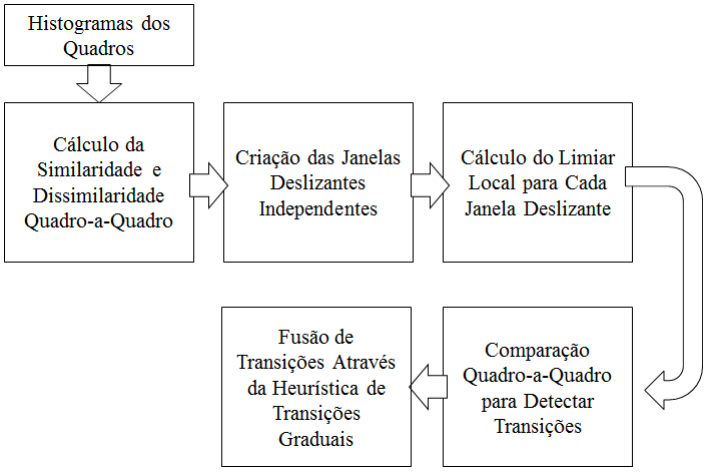


Figura 3.4: Fluxograma simplificado descrevendo os procedimentos realizados para a segmentação do vídeo em tomadas.

Inicialmente, estima-se a similaridade e a dissimilaridade quadro-a-quadro. A similaridade é medida através da intersecção de histogramas e a dissimilaridade através da distância euclidiana dos histogramas de quadros adjacentes, ambas apresentadas na Subseção 2.3.1.

A técnica de segmentação em tomadas é faz uso de duas janelas deslizantes independentes, uma baseada na intersecção e a outra baseada na distância euclidiana de histogramas. Tais métricas foram adotadas pois, segundo Doherty e Smeaton (2008), a intersecção de histogramas e a distância euclidiana apresentaram os melhores desempenhos para a detecção de eventos em imagens. Com isso, supõe-se que as possíveis deficiências de uma abordagem sejam amenizadas ou até mesmo superadas com o uso da outra abordagem.

O objetivo de cada janela deslizante é a de encontrar bordas de tomadas com alta dissimilaridade, ou seja, transições de tomadas facilmente reconhecíveis na análise de histogramas, também chamadas de transições "evidentes". Um exemplo de tal transição é apresentado na Figura 3.5.

A criação de cada janela deslizante é diferente entre a intersecção e a distância euclidiana. Na intersecção, encontra-se o menor valor de intersecção de histogramas entre todos os quadros adjacentes do vídeo, chamado de MIN. Com tal valor calculado, uma borda de janela é detectada quando:

(a) (b) (c) (d) (e) (f)

Figura 3.5: Exemplo de uma transição de tomadas "evidente", devido à alta dissimilaridade entre os histogramas dos quadros ilustrados em (c) e (d). Nesse exemplo, as figuras (a) até (c) e (d) até (e) formariam duas tomadas distintas.

- O valor da intersecção de histogramas for menor ou igual a 0.25 (25% de similaridade), ou;
- O valor da intersecção for menor ou igual a $MIN \cdot 1.5$ e, ao mesmo tempo, for menor que 0.4.

Para a distância euclidiana, encontra-se o maior valor da distância euclidiana entre todos os quadros adjacentes do vídeo, chamado de MAX. Com o valor calculado, uma borda de janela é detectada quando:

- O valor da distância euclidiana for maior ou igual 1.5 (75% de dissimilaridade), ou;
- O valor da distância euclidiana for maior ou igual a $MAX \cdot 0.85$ e, ao mesmo tempo, for maior que 0.9.

Destaca-se que os valores estabelecidos para a detecção das bordas de janelas foram determinados empiricamente. Tais testes seguiram o seguinte critério:

- Estimou-se um conjunto inicial de parâmetros, por exemplo, os limiares 0.2, 0.4, 0.6 e 0.8 para a intersecção de histogramas e os limiares 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6 e 1.8 para a distância euclidiana entre os histogramas.
- Com o conjunto inicial, calculou-se o parâmetro que obteve os melhores valores de precisão e abrangência médios entre todos os demais parâmetros do conjunto.
- Com o melhor parâmetro estimado, calcularam-se os valores de precisão e abrangência médios obtidos da vizinhança próxima, com incremento/decremento na ordem de 0.05 para cada teste.
- Caso os resultados apresentassem piora consecutivas (ou seja, ao aumentar o valor em 0.1), todos os valores maiores (caso de incremento) ou menores (caso de decremento) foram descartados.

- Por fim, o parâmetro escolhido é o valor que apresentou a maior taxa média de precisão e abrangência.

Com as bordas das janelas detectadas, o procedimento de segmentação em tomadas cria dois conjuntos independentes de janelas deslizantes que, em alguns casos, podem ter bordas compartilhadas. Nota-se que:

- Todas as bordas de janelas deslizantes são consideradas como transições de tomadas e são desconsideradas de análises posteriores.

- Todos os quadros pertencem à exatamente duas janelas, uma da intersecção de histogramas, outra da distância euclidiana de histogramas.

Com as transições evidentes detectadas, parte-se para a detecção de transições menos evidentes, ou seja, que apresentem variações menos perceptíveis em seus histogramas. Tais casos ocorrem principalmente em transições graduais de tomadas, onde o histograma de quadros adjacentes são semelhantes devido ao fato de apresentarem características tanto da tomada "anterior" como da tomada "posterior". Um exemplo de uma transição gradual de tomada que não seria detectada pelo procedimento anteriormente descrito é ilustrado na Figura 3.6

Para corrigir tais problemas, calcula-se um limiar local para cada uma das janelas deslizantes já detectadas. O cálculo de tal limiar, assim como das janelas deslizantes, é diferente entre a intersecção e a distância euclidiana de histogramas. Para a intersecção, o limiar é considerado como a metade do valor médio de intersecção na janela deslizante. No caso da distância euclidiana, o limiar local é igual a nove vezes o valor médio da distância euclidiana calculada na janela deslizante.

A Figura 3.7 exemplifica as janelas deslizantes e os limiares locais calculados em um conjunto de quadros de um vídeo de entrada.

Assim, com os limiares locais calculados, analisam-se linearmente todos os valores de intersecção e distância euclidiana calculados inicialmente. Os quadros que atingirem alguma das seguintes condições são ditos como "quadros de transição":

- O quadro analisado pertence a alguma borda de janela deslizante, seja da intersecção de histogramas, seja da distância euclidiana de histogramas.

- O quadro analisado possui valor de intersecção inferior ao limiar local da janela deslizante da intersecção de histogramas a que pertence.

- O quadro analisado possui valor de distância euclidiana superior ao limiar local da janela deslizante da distância euclidiana de histogramas a que pertence.

Figura 3.6: Exemplo de uma transição gradual do filme "A Beautiful Mind" que não seria detectada devido à alta similaridade dos histogramas entre quadros adjacentes.

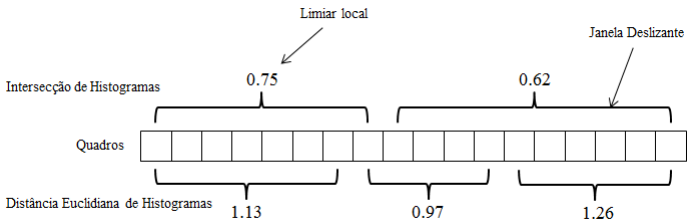


Figura 3.7: Exemplo da divisão dos quadros de entradas entre as janelas deslizantes e seus limiares locais calculados.

Com o processo realizado acima, detectam-se diversos quadros de transição capazes de segmentar o vídeo em tomadas. Porém, o algoritmo descrito até o momento apresenta uma *oversegmentation*, principalmente em relação às transições graduais, que, se detectadas, são compostas por diversos quadros de transição próximos. Assim, ao invés de uma única

transição gradual, o algoritmo poderia detectar dezenas de transições abruptas, afetando a precisão da técnica.

Logo, para detectar transições graduais, analisa-se cada quadro de transição detectado: caso a transição em análise estiver a quatro ou mais quadros de distância da transição analisada anteriormente e quatro ou mais quadros da próxima transição, ela é considerada como uma transição abrupta. Caso a distância seja inferior a quatro em algum dos casos, as transições são unidas, recursivamente, em uma só, formando uma transição gradual. A Figura 3.8 exemplifica o procedimento de união de transições para detectar transições graduais.

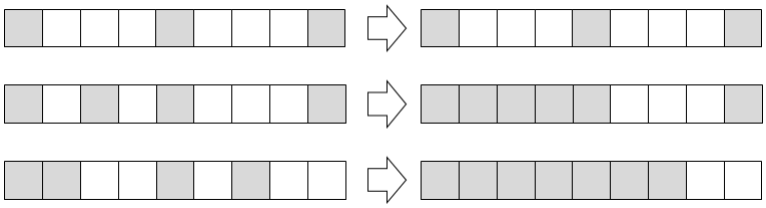


Figura 3.8: Exemplos de uniões de transições para a detecção de transições graduais, onde os quadros em cinza são quadros pertencentes à alguma transição. À esquerda, quadros antes da heurística de transições graduais e, à direita, quadros resultantes após a aplicação da heurística.

3.2 Fase Dois: Segmentação em Cenas

A segunda fase da técnica tem como entrada um fluxo de vídeo segmentado em tomadas, além de outras informações como os histogramas HSV de todos os quadros e os vetores de fluxo óptico de todos os quadros do vídeo. Destaca-se que a segmentação em cenas é modular, no sentido de que necessita como entrada as informações supracitadas, mas que o procedimento de segmentação em tomadas, como descrito na Subseção 3.1.3, poderia ser substituído por alguma outra técnica de segmentação em tomadas.

Inicialmente, na segmentação em cenas, encontra-se uma série de quadros-chave para representar cada uma das tomadas do vídeo. Após, calcula-se uma medida de coerência entre as tomadas do vídeo, seguida de uma detecção inicial das bordas das cenas. Infelizmente, nota-se que o resultado da análise da coerência visual:

- Retorna diversas cenas com curta duração.
- Diversas cenas são separadas em um conjunto de cenas pela presença de apenas uma tomada dissimilar.

- Cenas de ação, com baixa similaridade visual, geram um grande número de cenas adjacentes.

Assim, para reduzir a *oversegmentation*, uma série de procedimentos são realizados com o intuito de unir as diferentes cenas previamente detectadas, formando cenas com maior nível semântico através de medidas de dissimilaridade de histogramas, vetores de fluxo óptico e heurísticas simples.

3.2.1 Seleção dos Quadros-Chave

A seleção dos quadros-chave tem por objetivo encontrar um conjunto de quadros-chave para cada tomada capaz de representá-la adequadamente. A seleção de tais quadros é de suma importância para a segmentação em cenas baseadas em características visuais, pois, depois de selecionados, são os quadros-chave que serão utilizadas para comparar duas tomadas adjacentes e decidir se pertencem ou não à mesma cena. Com os quadros-chave, pode-se então realizar o processamento apenas sobre tais quadros, e não sobre a toda a tomada, resultando em um significativo aumento de desempenho.

Como citado na Subseção 2.5.1, selecionar apenas um quadro-chave pode resultar em uma representação inadequada da tomada. Assim, neste trabalho, desenvolveu-se uma técnica de seleção de quadros-chave capaz de representar tomadas simples, complexas e também casos onde houve erros no procedimento de segmentação do vídeo em tomadas. Para isso, ao invés de selecionar apenas um quadro-chave, a técnica desenvolvida procura encontrar um conjunto adequado de quadros-chave. A Figura 3.9 exemplifica a seleção de um conjunto de quadros-chave de um determinado segmento que apresenta duas tomadas.

Figura 3.9: Exemplo da seleção dos quadros-chave em um segmento de vídeo com duas tomadas distintas.

A seleção dos quadros-chave desenvolvida neste trabalho segue os seguintes procedimentos:

1. Calcula-se o grau de similaridade entre todos os quadros da tomada.
2. Seleciona-se o quadro que apresenta maior similaridade com os demais quadros.
3. Adiciona-se o quadro selecionado ao conjunto de quadros-chave, se nenhum quadro presente no conjunto de quadros-chave for similar ao quadro-chave candidato.
4. Repetem-se os passos 2 ao 4, até que o quadro candidato tenha baixa representatividade ou não exista quadro candidato.

O grau de similaridade é obtido através da intersecção de histogramas de todos os quadros formadores da tomada. Dois quadros são considerados "similares" caso a intersecção entre seus dois histogramas for igual ou maior que 95%. De mesma forma, dois quadros são ditos dissimilares caso o valor obtido seja inferior à 95%.

Neste trabalho, adotou-se 20% como representatividade mínima para que um quadro possa ser considerado como quadro candidato. Ou seja, um determinado quadro deve ser 95% "similar" a pelo menos 20% do número de quadros presente na tomada. Assim, essa técnica permite a seleção de diversos quadros-chave (ou apenas um quadro-chave) complementares que são suficientemente dissimilares entre si e que, ao mesmo tempo, são representativos aos demais quadros da tomada.

3.2.2 Cálculo da Coerência entre Tomadas

O cálculo da coerência entre tomadas é o núcleo central da técnica de segmentação em vídeo. Graças à medida de coerência entre tomadas, é possível agrupar um conjunto de tomadas em cenas.

A coerência de tomadas é uma medida de dissimilaridade, utilizada para medir o quão similar uma determinada tomada é frente a um conjunto determinado de tomadas anteriores. Neste trabalho, devido ao fato de fazer parte de uma técnica com baixo custo computacional e em estado-da-arte, adotou-se a medida de coerência de tomada, chamada de *SC* (*Shot Coherence*), definida por Rasheed e Shah (2003):

$$SC_j^i = \max_{f^x \in K_i, f^y \in K_j} \left(D\left(f^x, f^y\right) \right) \tag{3.4}$$

Onde f^x é o histograma de um quadro-chave do conjunto de quadros-chave K_i da tomada i, f^y é o histograma de um quadro-chave do conjunto de quadros-chave K_j da tomada j e $D\left(f^x, f^y\right)$ é a medida de comparação entre os histogramas f^x e f^y, no caso, a intersecção de histogramas. Assim, em outras palavras, o valor de coerência entre duas tomadas quaisquer é o maior valor da intersecção de histogramas entre quaisquer combinações dois-a-dois de quadros-chave das duas tomadas analisadas.

O cálculo da coerência da tomada, por si só, representa uma medida de coerência visual entre duas tomadas. Embora a segmentação em cenas pudesse ser realizada apenas com a análise do valor SC entre tomadas adjacentes, nota-se que o resultado seria influenciado por problemas como mínimos e máximos locais. Assim, é necessário um modo de analisar a variação da coerência de tomadas SC ao longo de algumas tomadas. Nesse sentido, Rasheed e Shah (2003) desenvolveram a métrica conhecida como BSC (*Backward Shot Coherence*), apresentada na Equação 3.5

$$BSC_i = \max_{1 \leq k \leq N} \left(SC_i^{i-k} \right) \tag{3.5}$$

Onde i é o índice da tomada, N é o tamanho da janela de busca adotado e SC_i^{i-k} é o valor de SC calculado entre a tomada i e a tomada $i - k$.

Após alguns testes terem sido realizados, notou-se que a medida BSC não considera a posição relativa de uma tomada em uma janela de busca. Assim, caso o tamanho de tal janela seja maior que a adequada, o valor BSC de uma tomada poderá ser determinado exclusivamente pelo valor SC em relação a uma tomada na extremidade oposta da janela de busca. Assim, o tamanho da janela tem uma influência muito grande no valor BSC, tornando a escolha de tal tamanho um fator determinante para a eficiência da técnica como um todo. Além disso, caso haja a presença de cenas com curta duração, as mesmas podem não ser corretamente detectadas, ocasionando falsos negativos.

Neste trabalho, para amenizar o impacto da decisão do tamanho da janela de busca, desenvolveu-se uma métrica diferente da especificada por Rasheed e Shah (2003). Para isso, introduziu-se um fator chamado de "Memória Temporal" (do Inglês *Temporal Memory* - TM), utilizado para aperfeiçoar o conceito da métrica de ser similar à memória humana. Com o valor TM, "pesos" diferentes são dados aos valores da coerência de tomadas de acordo com sua posição relativa (cronológica) em relação à tomada em análise. A nova métrica, com a introdução do fator TM, é chamada de *Backward Shot Coherence Weighted* (BSCW). A Equação 3.6 apresenta o cálculo do valor BSCW e a Equação 3.7 define o cálculo do valor TM

$$BSCW_i = \max_{1 \leq k \leq N} \left(SC_i^{i-k} \cdot TM_{N-k} \right) \tag{3.6}$$

Onde i é o índice da tomada, N é o tamanho da janela de busca adotado, SC_i^{i-k} é o valor de SC calculado entre a tomada i e a tomada $i - k$ e TM_{N-k} é o valor TM da tomada $N - k$.

$$TM_k = 1.0 + (0.05 \cdot k) \tag{3.7}$$

O valor TM_k, como definido na Equação 3.7, cresce conforme o índice k aumenta. Assim, o comportamento de seu valor na Equação 3.6 é a de incrementar o valor SC calculado em 5%, progressivamente, conforme a tomada sendo analisada aproxime-se da tomada ao qual se pretende calcular o valor BSCW. A Figura 3.10 exemplifica a diferença entre o valor BSC e BSCW para um dado conjunto de tomadas.

1	2	3	4	5	6	7	8	9	10	
0.64	0.56	0.42	0.61	0.13	0.24	0.38	0.40	0.43	0.44	

(a) Exemplo de cálculo do valor BSC da tomada em cinza.

1	2	3	4	5	6	7	8	9	10	
0.64	0.59	0.46	0.70	0.16	0.30	0.49	0.54	0.60	0.63	

(b) Exemplo de cálculo do valor BSCW da tomada em cinza.

Figura 3.10: Exemplo da diferença entre os valores de BSC e BSCW.

Na Figura 3.10(a), o valor de BSC da tomada em destaque seria determinado pela tomada de número um, ou seja, o valor BSC seria 0.64 (64% de coerência de tomada). Já a Figura 3.10(b) demonstra que o valor BSCW a ser escolhido seria determinado pela tomada quatro, ou seja, o valor BSCW seria de 0.70 (70% de coerência de tomada).

O valor BSCW, como definido, possui algumas propriedades, listadas a seguir:

- O valor da primeira tomada do vídeo é zero, devido à ausência de tomadas anteriores para o cálculo de algum valor SC.
- O valor cresce conforme tomadas similares vão sendo analisadas, ou seja, durante uma mesma cena, o valor BSCW tende a crescer.
- O valor apresenta uma queda ao se comparar uma tomada dissimilar às tomadas anteriores. Tal fato pode evidenciar uma transição de cena.
- O valor BSCW, embora se utilize de tomadas anteriores, tende a diminuir a importância das tomadas mais distantes da tomada em análise. Isso é devido ao valor de Memória Temporal (TM), que aumenta o "peso" do valor de tomadas mais próximas.

Por último, para o cálculo da coerência entre tomadas, é necessário definir o número máximo de tomadas que serão usadas para determinar o valor BSCW (o valor N na Equação 3.6). Um valor muito alto pode ocasionar a junção de cenas muito distantes em apenas uma, ignorando cenas menores entre ambas. Por outro lado, um valor consideravelmente baixo pode ocasionar um excesso de quedas do valor BSCW, já que a dinâmica de cenas

de média ou longa duração pode não ser completamente englobada em um tamanho de janelas extremamente restrito.

Em seu artigo, Rasheed e Shah (2003) utilizaram dez tomadas como tamanho de janelas para medir o valor BSC. Segundo os autores, o valor é adequado à base de vídeos utilizada em seus testes, que consistiam em trechos de filmes e *sitcoms*. Neste trabalho, também foi utilizado o valor definido por Rasheed e Shah (2003), ou seja, utilizam-se até dez tomadas anteriores como referência para o cálculo do valor BSCW para uma determinada tomada.

3.2.3 Criação de Bordas de Cenas

Após o cálculo do valor BSCW para todas as tomadas do vídeo pode-se, então, proceder com o agrupamento inicial do vídeo em cenas. Para isso, utiliza-se as propriedades do valor BSCW descritas na Subseção 3.2.2 para estimar um conjunto de cenas. A redução significativa do valor BSCW pode significar que aquela tomada não apresenta quadros-chave que foram considerados similares, usando a intersecção de histogramas, com qualquer outro quadro-chave de N tomadas anteriores. Com essa análise, é possível encontrar o começo de uma nova cena, já que mudanças nas texturas de fundos, nos objetos em destaque ou outros fatores resultaram em uma alta dissimilaridade com quadros-chave de tomadas anteriores.

Assim, uma borda de cena, ou seja, a transição de uma cena para outra é detectada quando:

- O valor BSCW sofre uma redução de 15% ou mais.
- O valor BSCW sofre uma redução de 5% ou mais, seguida imediatamente de outra redução em 5% ou mais.

Se alguma das condições anteriores for atendida, uma cena é criada unindo todas as tomadas até a próxima que também atender alguma das condições anteriores. Os valores de limiares foram estimados empiricamente através diversos testes em bases de vídeo do domínio "filmes".

Após a criação do conjunto inicial, o número de cenas é bastante elevado: diversas cenas são separadas pela presença de algumas tomadas com histogramas dissimilares às tomadas adjacentes. Assim, o número de falsos positivos obtido pela técnica é bastante elevado, resultando na chamada sobre segmentação (*oversegmentation*). Trabalhos semelhantes também encontraram tal problema (Bai et al., 2008; Rasheed e Shah, 2003), adotando, para isso heurísticas ou outros métodos posteriores para reduzir o número de falsos positivos.

Na técnica descrita neste trabalho, desenvolveu-se três métodos diferentes para reduzir a *oversegmentation*: a junção de cenas similares, a junção de cenas com dinâmica de cenas similares e, por fim, a junção de cenas intercaladas por cenas dissimilares. Cada um dos métodos é descrito nas subseções posteriores.

3.2.4 Junção de Cenas Similares

O primeiro método utilizado para a redução do número de falsos positivos, obtido pela união inicial de tomadas em cenas, é a junção de cenas adjacentes altamente similares. Tal excesso de falsos positivos usualmente ocorre, nesse caso, graças à presença de uma tomada bastante dissimilar dentre um conjunto de tomadas similares.

Como exemplo, sejam cinco tomadas, T_1, T_2, T_3, T_4 e T_5, cujos valores de BSCW são apresentados na Tabela 3.1

Tabela 3.1: Exemplo de cinco tomadas e seus valores BSCW.

Tomada	Valor BSCW
T_1	0.86
T_2	0.85
T_3	0.32
T_4	0.76
T_5	0.81

Assim, graças ao procedimento descrito na Subseção 3.2.3, serão criadas duas cenas C_1 e C_2, sendo C_1 formada pelas tomadas T_1 e T_2 e C_2 sendo formada pelas tomadas T_3, T_4 e T_5. Porém, nota-se que o valor BSCW das tomadas T_4 e T_5 são suficientemente altos para estarem na mesma cena que as tomadas T_1 e T_2. De fato, a cena C_2 só foi criada devido ao baixo valor BSCW apresentado pela tomada T_3. Nota-se que casos como esse ocorrem nos mais diversos tipos de domínios, como em telejornais (mudança de uma gravação no estúdio para uma gravação externa, por exemplo) e esportes (como a câmera colocando um jogador em *zoom*, por exemplo), além do próprio domínio de filmes. Um exemplo, no domínio filme, é ilustrado na Figura 3.11.

(a) (b) (c) (d) (e) (f)

Figura 3.11: Exemplo de um conjunto de tomadas, que formariam uma cena, mas que são divididas graças ao baixo valor BSCW na tomada ilustrada em (e).

Para o correto entendimento desta subseção, alguns conceitos e métricas utilizadas para os procedimentos de junção devem ser explicitados, entre os quais:

- **Quadros-chave de cenas** - É o conjunto formado pela união de todos os quadros-chave das tomadas que formam a cena.
- **Similaridade mínima de cenas** - É o menor valor da intersecção de histogramas entre todas as combinações de quadros-chave entre duas cenas quaisquer.
- **Similaridade média de cenas** - É o valor médio da intersecção de histogramas entre todas as combinações de quadros-chave entre duas cenas quaisquer.
- **Similaridade máxima de cenas** - É o maior valor da intersecção de histogramas entre todas as combinações de quadros-chave entre duas cenas quaisquer.

Assim desenvolveu-se um método de união de cenas que, embora tenham sido inicialmente divididas, apresentam um grau significativo de similaridade, sendo então fundidas em apenas uma cena. São adotadas três procedimentos para tal: a junção de cenas minúsculas, a junção de cenas similares adjacentes e a junção de cenas similares em janelas deslizantes.

O primeiro procedimento realizado, a junção de cenas minúsculas, parte do pressuposto que uma cena é usualmente representada por diversas tomadas consecutivas, ou seja, uma única tomada pode não representar corretamente toda a semântica embutida em uma cena.

A base do procedimento de junção de cenas minúsculas baseia-se na seleção das cenas constituídas por apenas uma tomada. Com isso, determina-se se a cena selecionada deve ou não ser fundida à cena anterior ou posterior. Em dois casos a cena minúscula não será unida a alguma cena adjacente nesse procedimento:

- Quando a tomada da cena minúscula tenha tamanho significativo, ou seja, tenha pelo menos 150 quadros de duração.
- Quando a similaridade média entre todos os quadros-chave da cena minúscula e suas cenas adjacentes for menor que 20%.

Pela primeira condição, uma única tomada grande o suficiente pode representar uma dinâmica complexa o suficiente para, ela própria, ser uma cena. Pela segunda condição, caso a cena minúscula for muito dissimilar as cenas adjacentes, então ela pode ser considerada como uma cena com apenas uma tomada.

Caso nenhuma das condições seja satisfeita, ou seja, a cena minúscula não tenha a duração ou seja dissimilar o suficiente, utiliza-se os valores de similaridade mínima, média e máxima para determinar se e com qual cena a cena minúscula deve ser fundida.

Caso o valor de similaridade média e máxima indicar, ambas, que a cena anterior ou posterior é a mais similar, a cena minúscula é unida com a cena que apresentar, ao mesmo tempo, a maior similaridade média e máxima.

Porém, caso a similaridade média e máxima não seja concordante, a cena minúscula será unida à cena anterior caso alguma das seguintes condições forem satisfeitas:

- A similaridade máxima seja pelo menos 20% maior que a similaridade máxima da cena minúscula com a cena posterior.
- A similaridade mínima seja pelo menos 10% maior que a similaridade mínima da cena minúscula com a cena posterior.

As mesmas condições aplicam-se para o caso da união da cena minúscula com a cena posterior, ou seja, a cena minúscula será unida com a cena posterior caso:

- A similaridade máxima seja pelo menos 20% maior que a similaridade máxima da cena minúscula com a cena anterior.
- A similaridade mínima seja pelo menos 10% maior que a similaridade mínima da cena minúscula com a cena anterior.

Caso não haja uma cena anterior, ou seja, a primeira cena do vídeo é uma cena minúscula, então se calcula a similaridade mínima com a cena posterior. Caso ela seja maior ou igual a 20%, a cena é fundida com a cena posterior. Caso a cena posterior não exista, ou seja, a última cena do vídeo é uma cena minúscula, então se a similaridade mínima com a cena anterior for de 20%, a cena é fundida com a cena anterior.

Com isso, a maioria das cenas minúsculas é fundida com cenas de maior tamanho graças à similaridade que apresentam entre si. Nota-se que nem todas as cenas minúsculas são fundidas: caso nenhuma condição seja satisfeita, a cena minúscula não será unida a nenhuma cena adjacente. Assim, embora a maioria das cenas minúsculas seja fundida às cenas mais similares adjacentes, ainda será possível existir cenas com apenas uma única tomada.

O segundo procedimento de junção de cenas similares não considera o tamanho da cena, apenas a sua similaridade. Assim, caso uma cena apresente alta similaridade com a cena imediatamente posterior, o procedimento avaliará se a junção entre as duas cenas deve ocorrer. Para isso, analisam-se os valores de similaridade mínima e máxima.

Dado duas cenas adjacentes, C_x e C_y, caso as duas cenas apresentem similaridade máxima de pelo menos 80%, ou similaridade máxima de pelo menos 60% e similaridade mínima de pelo menos 20%, elas podem ser unidas em apenas uma cena, já que são bastante similares.

Nota-se que, uma cena criada a partir da junção de duas cenas pelo segundo procedimento não será fundida novamente com qualquer outra cena neste procedimento. Tais casos serão tratados pelo último procedimento desta técnica.

Por fim, o terceiro e último procedimento do método de junção de cenas similares baseia-se, novamente, na similaridade entre cenas adjacentes. Porém, desta vez, o algoritmo utiliza-se de janelas deslizantes para fundir até três cenas em uma cena diretamente. Para isso, calculam-se os valores de similaridade máxima, média e mínima entre todas as combinações de quadros-chave de três cenas consecutivas. Sejam três cenas C_1, C_2 e C_3, assim fundem-se:

- As três cenas, caso a similaridade mínima entre a cena C_1 e C_3 for de pelo menos 50% e a similaridade máxima entre C_1 e C_3 for de pelo menos 80%. Caso o limiar de similaridade máximo não for atingido, então as cenas serão fundidas caso as cenas C_1 e C_3 apresentem similaridade média de 60% ou mais.

- Duas cenas adjacentes (C_1 e C_2 ou C_2 e C_3), caso a similaridade mínima for de pelo menos 25% e a similaridade máxima for de pelo menos 80%. Caso o limiar de similaridade máximo não for atingido, então as cenas serão fundidas caso apresentem similaridade média de 60% ou mais.

Assim, o último procedimento procura unir três cenas em apenas uma, desconsiderando os valores de similaridade da cena mediana. Nota-se que, nesse caso, o limiar de similaridade mínimo exigido é de 50%, mais rígido que no caso da união de duas cenas. Na prática, em poucas ocasiões o procedimento realizará a junção de três cenas em apenas uma. Mesmo assim, quando tal junção ocorre, o número de falsos positivos é reduzido significativamente.

3.2.5 Junção de Cenas com Dinâmica de Movimento Similar

Após a execução do procedimento descrito na Subseção 3.2.4, a maioria dos casos onde uma tomada visualmente dissimilar resultou na divisão de uma cena única em duas cenas são resolvidos. Porém, nota-se que o procedimento parte do pressuposto de que há, entre os segmentos a serem unidos, um grau significativo de similaridade visual. Porém, em cenas mais complexas, a progressão da ação pode incluir tomadas significativamente dissimilares uma das outras.

Rasheed e Shah (2003), por exemplo, destacam em seu trabalho que o uso exclusivo da coerência de tomadas resultou em um elevado número de falsos positivos (sobre segmentação) em cenas de ação, que apresentam alta movimentação e tomadas com baixa similaridade visual. Para enfrentar tal problema, eles desenvolveram um procedimento

baseado na extração dos vetores de movimento de um fluxo de vídeo de entrada no formato MPEG-1. Com os vetores de movimento obtidos, Rasheed e Shah (2003) estimam a "quantidade de movimento" entre duas cenas adjacentes e, caso apresentem um alto grau de movimentação, as cenas são unidas em uma única cena.

Neste trabalho, desenvolveu-se um método cujo objetivo é a junção de cenas adjacentes com dinâmica de movimento similar. Nesse método, considera-se os vetores de fluxo óptico calculados na primeira fase do algoritmo e descritos na Subseção 3.1.2.

No procedimento de cálculo dos vetores do fluxo óptico, obtém-se um conjunto de vetores para cada quadro do vídeo. Dado tais vetores, elimina-se os vetores cujo tamanho calculado pela distância euclidiana entre dois pontos seja menor que 1.0. Esses vetores minúsculos representam, na sua maioria, pequenas oscilações de *pixels* tais como folhas de árvores se mexendo ou outros movimentos sutis e pouco significativos.

Depois de eliminados, conta-se o número de vetores restantes para cada quadro. Com isso, pode-se calcular o valor chamado de $FluxoTomada$, definido na Equação 3.8:

$$FluxoTomada_i = \sum_{a=0}^{N} FluxoQuadro_a \tag{3.8}$$

Onde N é o número de quadros que a i-tomada possui. Com o $FluxoTomada$ calculado, pode-se então calcular o valor $FluxoCena$, definido na Equação 3.9:

$$FluxoCena_i = \frac{\sum_{a=0}^{N} FluxoTomada_a}{N} \tag{3.9}$$

Onde N, nesse caso, é o número de tomadas da i-cena.

Em outras palavras, o $FluxoCena$ contém a média da quantidade de vetores de movimento das tomadas que a compõe. Caso uma tomada tenha alto número de vetores de movimento, as demais tomadas da cena podem reduzir o impacto na média, caso apresentem reduzido número de vetores de movimento.

Com o valor de $FluxoCena$ calculado para todas as cenas do vídeo, parte-se para a junção das cenas caso 1) apresentem $FluxoCena$ similares e/ou 2) tenham valor de $FluxoCena$ acima de determinado limiar.

Para o primeiro caso, analisam-se as cenas em janelas de busca com tamanho três. Dada três cenas C_1, C_2 e C_3:

- Unem-se as cenas C_1, C_2 e C_3 caso a diferença entre o valor de $FluxoCena$ entre C_1 e C_3 seja menor ou igual a 25%. Além disso, o valor de $FluxoCena$ entre as cenas C_1 e C_2 ou o valor entre as cenas C_2 e C_3 deve ser inferior ou igual à 25%.

- Unem-se as cenas C_2 e C_3 caso a diferença entre o valor de $FluxoCena$ entre C_2 e C_3 seja menor ou igual a 25%.

- Unem-se as cenas C_1, C_2 caso a diferença entre o valor de $FluxoCena$ entre C_1 e C_2 seja menor ou igual a 25%.

Com isso, duas ou mais cenas adjacentes com valor de $FluxoCena$ similares são unidas. Esse procedimento tem por objetivo unir cenas com dinâmica similar de movimento, formando cenas com uma movimentação similar. Por exemplo, cenas consecutivas de uma entrevista podem ser unidas em apenas uma única cena devido à baixa movimentação presente em todas as cenas.

O segundo caso, a junção de cenas com alto valor de $FluxoCena$, é um caso também explorado por trabalhos similares. Rasheed e Shah (2003), por exemplo, amenizam a sobre segmentação em cenas com alta movimentação através da análise da "quantidade de movimento" obtido dos vetores de movimento do fluxo de entrada. Segundo eles, alguns tipos de cenas, como de ação, apresentam como característica o pequeno tamanho das cenas e a alta movimentação. Graças a isso, a análise de coerência de tomadas não é capaz de identificar esse conjunto de cenas como apenas uma única cena, resultando em uma sobre segmentação devido às cenas com alta movimentação.

Assim, nesse trabalho, utilizou-se o fluxo óptico para, também, estimar a quantidade de movimento de uma cena. Caso o valor de $FluxoCena$ seja elevado, significa que a cena possui tomadas formadas por quadros com diversos vetores de movimento significativos, já que vetores com tamanho minúsculo foram removidos previamente. Assim, cenas que apresentem tal comportamento podem ser ditas como cenas de "alta" movimentação.

Para unir cenas que apresentem alta movimentação, é necessário encontrar o valor adequado para o limiar. Caso o limiar seja inferior ao ideal, cenas com pequena movimentação podem ser unidas inapropriadamente, e, caso o valor seja acima do ideal, cenas com razoável movimentação podem ser completamente ignoradas.

Nesse sentido, criou-se uma forma adaptável de calcular o limiar de fluxo óptico de um vídeo. O cálculo do limiar, chamado de *Limiar Fluxo Optico*, é descrito na Equação 3.10:

$$Limiar\ Fluxo\ Optico = \frac{Video_{largura} \cdot Video_{altura}}{Tamanho\ Janela\ Fluxo\ Optico^3} \tag{3.10}$$

Onde $Tamanho\ Janela\ Fluxo\ Optico$ é o tamanho da janela de busca adotado no cálculo dos vetores do fluxo óptico.

Com isso, o valor de limiar varia conforme a resolução do vídeo e o tamanho da janela de busca dos fluxos ópticos. Tal cálculo é justificado, pois, ao aumentar a resolução do vídeo, mais *pixels* podem ser considerados através do fluxo óptico. Além disso, com o tamanho da janela de busca maior, mais *pixels* podem ser "casados" na busca dos vetores de fluxo óptico, gerando mais vetores em comparação com outra procura com tamanho de janela de busca menor.

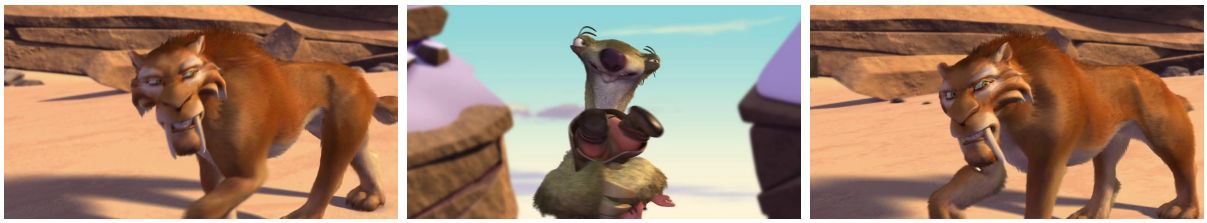
Assim, procura-se linearmente as cenas que possuam $FluxoCena$ maiores que o limiar calculado $Limiar\,Fluxo\,Optico$. Se o valor for superior ao limiar e a cena for adjacente à outra cena cujo valor de $FluxoCena$ também seja superior ao limiar, então as duas cenas são fundidas em uma só.

Caso três cenas consecutivas tenham seus valores de $FluxoCena$ maiores que o limiar, as duas cenas adjacentes com o maior valor serão fundidas. Porém, como o método é executando enquanto houve alguma junção, possivelmente a cena gerada através da junção também se fundirá com a cena adjacente a ela, já que a junção de duas cenas com valores de $FluxoCena$ acima do limiar resulta em uma cena com valor $FluxoCena$ também superior ao limiar.

Ao final desse método, a maioria dos falsos positivos foi eliminada. Cenas são formadas por tomadas adjacentes com quadros-chave altamente similares e/ou que possuam dinâmica de movimento similar ou muito alta.

3.2.6 Junção de Cenas Intercaladas por Cenas Dissimilares

Mesmo com a adoção do procedimento de junção de cenas similares, apresentado na Subseção 3.2.4, e com dinâmica de movimento similares apresentado na Subseção 3.2.5, alguns casos de falsos positivos persistem. A maioria desses casos relaciona-se com cenas que apresentam dinâmica de movimento dissimilar e cenas adjacentes com baixa similaridade em termos de histogramas. Um exemplo desse tipo de falso positivo é ilustrado na Figura 3.12.



(a) Quadro-chave da Cena 1 (b) Quadro-chave da Cena 2 (c) Quadro-chave da Cena 3

Figura 3.12: Exemplos de falsos positivos que persistem mesmo após a aplicação de métodos para junção de cenas similares e/ou com movimentação similar.

Ou seja, como demonstrado no exemplo da Figura 3.12, cenas que são intercaladas por uma cena "diferente" das adjacentes fazem com que três cenas sejam detectadas, ao invés de apenas uma. Analisando-se o contexto, em tais situações a câmera muda (indicando mudança de tomada), mas o assunto não (indicando que a cena é a mesma). Embora tais cenas devessem ter sido unidas no terceiro procedimento descrito na Subseção 3.2.4, o fato de que as cenas são novamente fundidas no procedimento de junção de cenas com

dinâmica de movimento similar (descrito na Subseção 3.2.5), faz com que novas cenas intercaladas surjam.

Para casos assim, desenvolveu-se um método de junção de cenas, baseada na intersecção de histogramas, capaz de reconhecer e encontrar cenas intercaladas por uma cena muito dissimilar. Sejam três cenas, C_1, C_2 e C_3, tais que a cena C_2 seja muito dissimilar as cenas C_1 e C_3. Assim, serão unidas as três cenas em apenas uma se todas as condições abaixo forem atingidas:

- C_1 e C_3 apresentem similaridade máxima de pelo menos 95%, ou seja, houver pelo menos um par de quadros-chave com 95% de similaridade.
- C_1 e C_3 apresentem similaridade média de pelo menos 50%, ou seja, a média da intersecção dos histogramas dos quadros-chave tiver similaridade maior ou igual a 50%.
- C_1 e C_3 apresentem similaridade mínima de pelo menos 25%, ou seja, para quaisquer dois pares de histogramas dos quadros-chave, a intersecção entre as mesmas deve ser maior ou igual a 25% de similaridade.

Embora seja possível encontrar e reduzir um grande número de falsos positivos, na prática poucas cenas atendem a tais limiares. Como vantagem, porém, destaca-se que cenas unidas com tal método apresentam baixo índice de erro, ou seja, poucos falsos positivos ou falsos negativos são criados com a aplicação dessa técnica.

Ao final deste procedimento, a técnica retorna um conjunto de transições de cenas, criadas a partir da junção progressiva das tomadas detectadas na primeira fase do algoritmo. Uma avaliação em termos de precisão, abrangência e F_1 da segmentação em tomadas e em cenas é descrita no Capítulo 4.

3.3 Detalhes da Implementação da Técnica

Implementou-se a técnica descrita nesse trabalho utilizando-se a linguagem Java na versão 1.7. Para auxiliar o desenvolvimento, utilizou-se a biblioteca JavaCV[3], uma interface em Java para bibliotecas como a OpenCV e FFMpeg. A biblioteca OpenCV, por sua vez, é formada por mais de 2500 algoritmos clássicos e em estado-da-arte na área de visão computacional e aprendizagem de máquina, criados especialmente para acelerar o desenvolvimento de aplicações nessas áreas. Já a biblioteca FFMpeg é especializada em prover uma ampla gama de codificadores e decodificadores otimizados para diversos formatos de vídeo e áudio.

[3]http://code.google.com/p/javacv/

Diversas funcionalidades foram implementadas através do uso de recursos da biblioteca JavaCV, incluindo:

- Decodificação do vídeo de entrada.
- Conversão do espaço de cor RGB para o espaço de cor HSV e para a escala-de-cinza.
- Extração dos histogramas.
- Cálculo da intersecção de histogramas.
- Detecção de *pixels* de bordas da imagem dos quadros.
- Cálculo dos vetores de fluxo óptico.

Embora apresente diversos recursos disponíveis aos programadores, algumas funcionalidades tiveram de ser implementadas manualmente, tais como:

- Cálculo da distância euclidiana entre dois histogramas.
- Cálculo dos valores SC e $BSCW$.
- Cálculo dos diversos limiares utilizados.

Para aproveitar-se dos processadores *multi-core*, a técnica implementada utilizou-se de uma abordagem *multi-thread* capaz de reduzir o tempo total de processamento.

Na primeira fase da técnica, a segmentação em tomadas, o processo de decodificação dos quadros é realizada pela FFMpeg, utilizando-se todos os núcleos disponíveis do processador. Após um determinado número de quadros terem sido decodificados e armazenados em memória, entra em cena o processo de extração dos histogramas e o cálculo do fluxo óptico. Como tais procedimentos utilizam-se dos quadros decodificados mas não são dependentes entre si, tais procedimentos são realizados concorrentemente. Após ambos terem processados todos os quadros previamente decodificados, os quadros são liberados ou removidos da memória e o processo se reinicia até que todo o vídeo tenha sido decodificado e processado. Ao final, o procedimento de segmentação em tomadas é realizado utilizando-se apenas uma *thread* para seu processamento.

O número de quadros armazenados na memória principal influencia diretamente o custo de memória principal da aplicação desenvolvida. Como o quadro armazenado é decodificado por padrão no espaço de cor RGB sem quantização, cada *pixel* do quadro necessita de três *bytes* de memória. Caso o quadro possua resolução 1920x1080, por exemplo, são necessários 5.93MB de memória para armazená-lo. Assim, embora um alto

número de quadros armazenados e disponíveis para processamento resulte em maior velocidade de processamento, um número excessivo pode tornar o aplicativo inoperável ou reduzir drasticamente o desempenho em computadores com memória insuficiente.

Para a segunda fase do algoritmo, a maioria dos procedimentos é realizada de maneira convencional, sem o uso de *threads* concorrentes. Porém, os métodos de seleção dos quadros-chave e o cálculo do valor de coerência entre tomadas são realizados em mais de uma *thread*. Nota-se que o cálculo da coerência de tomadas só pode ser iniciado após todos os quadros-chave terem sido selecionados, já que o valor BSCW depende dos mesmos.

Em termos de saída da aplicação, o algoritmo retorna ao usuário dois arquivos de registro, um para a segmentação em tomadas e outro para a segmentação em cenas. A segmentação em tomadas retorna o número do quadro inicial e final que pertencem às transições de tomadas, que podem ser de dois tipos: abruptas e graduais. O resultado é armazenado em um arquivo no padrão XML, baseado em um formato utilizado pela TRECVid[4]. Um arquivo de registro de exemplo da segmentação em tomadas do vídeo "video.mp4" com 6543 quadros de duração é apresentado na Figura 3.13.

```
<?xml version="1.0" encoding="UTF-8"?>
<refSeg src="video.mp4" creationMethod="Manual" totalFNum="6543">
        <trans type="CUT" preFNum="814" postFNum="815" />
        <trans type="CUT" preFNum="2497" postFNum="2498" />
        <trans type="CUT" preFNum="4251" postFNum="4252" />
        <trans type="GRA" preFNum="4293" postFNum="4299" />
        <trans type="CUT" preFNum="4354" postFNum="4355" />
        <trans type="GRA" preFNum="5996" postFNum="6010" />
        <trans type="GRA" preFNum="6046" postFNum="6054" />
        <trans type="CUT" preFNum="6137" postFNum="6138" />
</refSeg>
```

Figura 3.13: Arquivo de registro de exemplo para a segmentação em tomadas do arquivo "video.mp4".

Nota-se que, ao contrário do modelo originalmente proposto, o tipo de transição para a segmentação em tomadas é do tipo "CUT" para transições abruptas e "GRA" para transições graduais. Tal diferenciação ao modelo original, que diferenciava os diversos tipos de transições graduais, foi adotada pois os filmes apresentam efeitos complexos cuja classificação poderiam causar ambiguidades e, ainda, devido ao fato da técnica de segmentação de vídeo não diferenciar entre os diferentes tipos de transições graduais.

Para a segmentação em cenas, a aplicação retorna outro arquivo de registro XML semelhante ao retornado na segmentação em tomadas. Porém, a transição retornada é do tipo "SCE", representando uma transição de cena. Nesse caso, não se diferencia caso a transição tenha duração de apenas um quadro (como uma transição abrupta) ou uma transição entre vários quadros (como uma transição gradual).

[4]http://goo.gl/NUjPG

Uma descrição mais detalhada da implementação da técnica pode ser encontrada em (Trojahn e Goularte, 2013), onde se descreve a divisão da implementação em classes, os métodos disponibilizados por cada classe e outros detalhes de implementação da técnica descrita neste trabalho.

3.4 Considerações finais

Neste capítulo apresentou-se a técnica de segmentação de vídeo em cenas desenvolvida. A Seção 3.1 descreveu a primeira fase do algoritmo, a segmentação em tomadas, essencial para a segunda fase do algoritmo, a segmentação em cenas, descrita na Seção 3.2. Por fim, na Seção 3.3, descreveu-se uma implementação da técnica desenvolvida.

A técnica de segmentação descrita realiza a segmentação de forma progressiva: partindo-se dos quadros do vídeo, agrupa-os em tomadas e, então, agrupa as tomadas progressivamente até formar um conjunto de cenas. Embora a técnica seja baseada em histogramas (segmentação em tomadas, seleção de quadros-chave, análise de coerência visual) e no fluxo óptico (análise da dinâmica de movimento), a presença dos diversos procedimentos heurísticos propostos, como a análise de potenciais cenas adjacentes, auxilia a técnica em si a alcançar um conjunto de cenas com maior grau semântico que trabalhos relacionados baseados apenas em coerência visual, por exemplo.

A divisão da técnica em duas fases distintas, a segmentação em tomadas e a segmentação em cenas, pode auxiliar trabalhos futuros no sentido de facilitar que melhorias possam ser feitas em cada fase individualmente. Por exemplo, na fase de segmentação em tomadas, a introdução de uma análise para a detecção de *flashes*, uma das situações onde os histogramas apresentam problemas, poderia melhorar a eficácia da detecção de tomadas e até mesmo a de cenas, sem precisar modificar a segunda fase do algoritmo.

Como citado anteriormente, a segmentação em tomadas é realizada através da análise dos histogramas dos quadros do vídeo. A utilização de histogramas, uma técnica adotada em diversos outros trabalhos, garante um baixo custo computacional aliada a boas taxas de precisão e abrangência na segmentação em tomadas, como descrito no Capítulo 4.

A seguir, apresentam-se algumas considerações importantes acerca da segunda fase do algoritmo, a segmentação em cenas:

- O procedimento de seleção de quadro-chave foi desenvolvido de forma a tentar representar adequadamente tomadas complexas (com conteúdo visual altamente dinâmico) ou onde ocorreram erros na segmentação em tomadas (por exemplo, em transições graduais entre tomadas). A seleção de múltiplos quadros-chave, porém, pode não ser adequada em tomadas simples, com baixa dinâmica visual, e onde não ocorreram erros na detecção de tomadas. Em tais casos, o procedimento de com-

paração exaustiva impacta negativamente no custo computacional da técnica sem, possivelmente, prover qualquer benefício para as fases posteriores, haja vista que um único quadro-chave poderia representar a tomada corretamente.

- O número de tomadas a ser analisado para a coerência visual, o parâmetro principal para o cálculo do valor BSCW, pode variar conforme o domínio de entrada do vídeo e outras características intrínsecas do vídeo, tais como a duração das tomadas. Em telejornais, por exemplo, a transição de cena de uma notícia de baixa duração, com poucas tomadas externas, poderia ser completamente desconsiderada devido à similaridade nas tomadas onde o âncora está presente. Infelizmente, não foi encontrado um método adequado para a determinação do tamanho da janela de análise ideal, sendo por isso adotado o mesmo valor reportado em trabalhos relacionados.

- Os procedimentos de junção de cenas similares e de cenas similares intercaladas por cenas dissimilares visam a ampliar a semântica incluída em cada segmento detectado pela técnica. Tais procedimentos foram desenvolvidos heuristicamente ao analisar os resultados obtidos pela segmentação baseada apenas na coerência visual. Assim, tais procedimentos são fortemente baseados na definição de cena apresentada na Seção 2.2, sendo, por isso, os métodos que possivelmente resultam no maior ganho semântico da técnica em si.

Além disso, é importante citar que diversos parâmetros e valores utilizados para a técnica de segmentação foram determinados empiricamente através de diversos testes exaustivos, com o objetivo de otimizar os valores de precisão e abrangência obtidos pela técnica com a base de vídeos especificada no Capítulo 4.

Por fim, uma implementação da técnica foi descrita brevemente na Seção 3.3, apresentando alguns detalhes de implementação tais como a decodificação e a extração *multi-thread* dos histogramas e do fluxo óptico através da biblioteca OpenCV na linguagem Java.

Capítulo

4

Avaliação da Técnica

Este capítulo descreve as avaliações realizadas sobre a técnica descrita no Capítulo 3. A Seção 4.1, apresenta a base confiável utilizada para os testes realizados sobre a implementação da técnica. Nas Seções 4.2 e 4.3 são apresentados, respectivamente, os resultados dos testes de desempenho e de eficácia da técnica desenvolvida ao segmentar os vídeos de entrada da base de dados confiável.

Na Seção 4.4 é apresentada uma comparação dos resultados obtidos pela técnica desenvolvida com outras técnicas de segmentação em cenas em estado-da-arte. A Seção 4.5 apresenta algumas limitações encontradas na técnica desenvolvida. Por fim, na Seção 4.6, são apresentadas algumas considerações finais sobre este capítulo.

4.1 Descrição da Base Confiável

A base confiável é um conjunto de vídeos segmentados manualmente de maneira a sintetizar a saída esperada de um algoritmo que não apresente nenhum falso positivo ou falso negativo, ou seja, que possua 100% de precisão e abrangência.

No caso de segmentação de vídeo em tomadas e cenas, uma base confiável consiste em algum tipo de arquivo de registro contendo informações necessárias para identificar as tomadas e/ou cenas e o conjunto de vídeos correspondente. Com tal registro, compara-se a base confiável com o arquivo de registro obtido por uma técnica automática de segmen-

tação, possibilitando o cálculo dos valores de precisão, abrangência e F_1 (Chasanis et al., 2009; Chen e Li, 2010; Rasheed e Shah, 2003).

Idealmente, diversas técnicas de segmentação utilizariam a mesma base confiável para que possam ser comparadas diretamente. Infelizmente, não foi encontrada uma base de vídeos adequada para a segmentação em cenas. Por exemplo, um importante repositório de bases confiáveis de vídeo, o TRECVid[1], apresenta os seguintes problemas:

- Grande parte dos vídeos disponibilizados possuem curta duração tornando-se, assim, inadequados para uma análise de segmentação em cenas, haja vista que tais vídeos possuem poucas (ou até mesmo nenhuma) transições de cenas.

- Alguns vídeos não estão disponíveis gratuitamente.

- Não há arquivos de registro disponibilizados representando uma base confiável de vídeos para a segmentação em cenas.

Assim, devido à ausência de uma base confiável que pudesse ser utilizada facilmente, criou-se uma base confiável própria contendo vídeos do domínio filmes. Tal domínio foi escolhido devido ao fato de que tais vídeos podem ser facilmente obtidos a baixo custo, como em locadoras de vídeo e assemelhados, além de apresentarem grande duração e possuírem diversas cenas.

Foram selecionados cinco filmes de diferentes gêneros e anos de produção para diversificar a base de vídeos. O nome dos filmes escolhidos, o ano de produção, o gênero (baseado nas informações do IMDB[2]) e a duração do filme são apresentadas na Tabela 4.1.

Tabela 4.1: Detalhes dos filmes escolhidos para formar a base confiável.

Nome	**Ano de produção**	**Gênero**	**Duração**
A Beautiful Mind	2001	Biografia, Drama	135m
Back to the Future	1985	Aventura, Comédia, Ficção Científica	116m
Gone in Sixty Seconds	2000	Ação, Crime, *Thriller*	118m
Ice Age	2002	Animação, Aventura, Comédia	81m
Pirates of the Caribbean: The Curse of the Black Pearl	2003	Ação, Aventura, Fantasia	143m

[1]http://trecvid.nist.gov/trecvid.data.html
[2]http://www.imdb.com/

Os vídeos foram digitalizados e codificados utilizando-se o codificador x264[3] para o padrão H.264/AVC na taxa de 24 quadros por segundo (fps). Foram removidos todos os trechos iniciais como apresentação dos estúdios de produção e os créditos iniciais, além dos créditos finais do filme. Tais trechos apresentam baixa relação semântica em relação ao vídeo e, por isso, podem ser desconsiderados. Os detalhes específicos da base de vídeos, tais como a resolução e número do quadro inicial/final do vídeo, são apresentados na Tabela 4.2.

Tabela 4.2: Detalhes de codificação dos vídeos escolhidos para a base confiável.

Nome	**Resolução**	**Nº quadros**	**Nº quadro inicial**	**Nº quadro final**
A Beautiful Mind	1280x688	182870	2152	157914
Back to the Future	1280x696	159929	900	160829
Gone in Sixty Seconds	1920x816	155762	4836	160598
Ice Age	1280x694	107781	685	108466
Pirates of the Caribbean: The Curse of the Black Pearl	1280x528	191152	396	191548

Da base confiável, extraiu-se ainda um trecho de cada vídeo onde a técnica obtém melhores resultados de segmentação em termos de precisão, abrangência e F_1. Tais trechos foram extraídos, mantendo-se o codificador do fluxo original, com o intuito de compará-los com outras técnicas de segmentação de vídeo, como a técnica desenvolvida por Rasheed e Shah (2003), que usam pequenos trechos do filme original para a análise de precisão e abrangência. Os detalhes de cada trecho extraído como quadro inicial e quadro final (considerando o vídeo original, sem a remoção dos créditos inicial/final), número de quadros total e duração são apresentadas na Tabela 4.3.

Com a base de vídeos estabelecida, realizou-se a segmentação manual dos vídeos em tomadas e cenas. A segmentação em tomadas foi realizada quadro-a-quadro obtendo-se dois tipos de transição de tomadas: abruptas e graduais. Anotaram-se os números do quadro inicial e final da transição entre as tomadas, obtendo-se um registro completo entre todas as transições de tomadas do vídeo.

As transições de tomadas abruptas foram segmentadas analisando-se diferenças entre dois quadros adjacentes. Caso uma mudança brusca na posição dos objetos, do ângulo da câmera ou do próprio cenário é percebida, uma transição abrupta é sinalizada. A Figura 3.5 exemplifica uma transição abrupta.

[3]http://www.videolan.org/developers/x264.html

Tabela 4.3: Detalhes dos trechos de vídeos extraídos dos filmes que compõe a base confiável.

Nome	Quadro inicial	Quadro final	Nº quadros	Duração
T. A Beautiful Mind	21747	74895	53148	36m e 57s
T. Back to the Future	50695	97077	46382	32m e 15s
T. Gone in Sixty Seconds	10790	93885	83095	57m e 46s
T. Ice Age	6702	63138	56436	39m e 14s
T. Pirates of the Caribbean: The Curse of the Black Pearl	396	42582	42186	29m e 20s

Para a transição gradual, considerou-se como inicio da transição o quadro onde pode ser percebido o começo do efeito de transição, o final da transição corresponde ao primeiro quadro onde não é mais perceptível o efeito de transição. Por exemplo, na Figura 3.6, o 6º quadro (da esquerda para a direita, de cima para baixo) é considerado como início da transição, sendo o 44º quadro o final da transição, haja vista que o antebraço do personagem não é mais distinguível.

A segmentação manual em cenas considerou os conceitos apresentados na Seção 2.2 para a obtenção dos limiares de transição entre as cenas. Nota-se que as transições obtidas coincidem com as transições de tomada correspondentes, ou seja, as transições obtidas para a segmentação em cenas é um subconjunto das transições obtidas para a segmentação em tomadas.

Os resultados da segmentação manual realizada são apresentados, em termos de número de transições de tomadas e cenas, para os vídeos completos e seus trechos, na Tabela 4.4.

A base escolhida possui grande variedade, como demonstrado na Tabela 4.4. Nota-se que o filme *Pirates of the Caribbean: The Curse of the Black Pearl* é o que apresenta o maior número de transições de cenas, enquanto que o filme *Back to the Future* o menor. Considerando a proporção de transições de tomadas e as transições de cenas, o filme *Ice Age* apresenta as cenas com menor duração, resultando em um elevado número de transições de cenas. Por exemplo, a cena do filme completo *Ice Age* que começa no tempo 03:22 e termina no tempo 03:38, contendo apenas duas tomadas.

Nota-se ainda que, embora apresentem duração total semelhante, os filmes *Back to the Future* e *Gone in Sixty Seconds* possuem uma diferença significativa no número de transições de tomadas e cenas. Isso se deve ao fato de que o primeiro filme, do estilo Drama e Biografia, apresenta tomadas maiores com diálogos, resultando em cenas com

Tabela 4.4: Número de transições de tomadas e cenas, obtidas na segmentação manual, dos vídeos completos e seus trechos.

Nome	**Nº de transições de tomadas**	**Nº de transições de cenas**
A Beautiful Mind	1590	96
Back to the Future	1350	81
Gone in Sixty Seconds	2681	159
Ice Age	1365	133
Pirates of the Caribbean: The Curse of the Black Pearl	2659	168
T. A Beautiful Mind	433	34
T. Back to the Future	330	20
T. Gone in Sixty Seconds	1200	97
T. Ice Age	6702	73
T. Pirates of the Caribbean: The Curse of the Black Pearl	609	39

maior duração. No segundo caso, os trechos de ação aumentam significativamente o número de transições de tomadas e, consequentemente, o número de cenas, resultando na disparidade entre o número de transições entre tais filmes. Em suma, o número de transições de cenas varia conforme o estilo de produção, o gênero e outras características intrínsecas do filme.

4.2 Testes de Desempenho

Os testes de desempenho visam aferir a velocidade de processamento da técnica descrita neste trabalho. Com isso, pode-se, por exemplo, definir cenários de utilização mais apropriados para a técnica, principalmente em casos onde a latência é um fator crítico.

Para medir o desempenho, adotou-se a metodologia de medir o tempo total de processamento da técnica. O tempo total de processamento, neste trabalho, foi estimado com o auxílio da IDE NetBeans[4], que indica o tempo total gasto com o processo em si. Para isso

[4]http://netbeans.org/

utilizou-se um computador pessoal equipado com um processador Intel® Core™ i7 980[5] com 16GB de memória principal rodando a implementação descrita na Seção 3.3.

Com o valor de tempo total gasto para executar o algoritmo, analisando-se o número de quadros do vídeo de entrada, obtém-se a taxa de processamento, definida na Equação 4.1.

$$Taxa\,de\,Processamento = \frac{Numero\,de\,Quadros\,do\,Video}{Tempo\,de\,Processamento} \tag{4.1}$$

Onde *Taxa de Processamento* é dado em quadros por segundo (fps) e *Tempo de Processamento* é dado em segundos (s). Com tal taxa, pode-se analisar o quão próximo o algoritmo situa-se para operações ao vivo, em comparação com os padrões de TV Digital pelo mundo, por exemplo.

Uma técnica possui uma alta velocidade de processamento caso alcance uma alta *Taxa de Processamento*. Tal métrica indica quantos quadros o algoritmo é capaz de processar por segundo e, caso tal taxa supere a taxa de exibição do filme (24 fps para os vídeos da base de testes utilizada), a técnica de segmentação pode ser utilizada concorrentemente com a exibição do vídeo, por exemplo.

Os testes de desempenho foram realizados executando-se todos os procedimentos das duas fases da técnica com todos os vídeos da base confiável, anotando-se o tempo total de processamento e, assim, obtendo a taxa de processamento em FPS.

Nota-se que o desempenho da técnica ao realizar a segmentação dos trechos dos fluxos de vídeo é idêntico aos resultados obtidos ao se segmentar o fluxo completo. Ou seja, a taxa de processamento (em FPS) média é similar entre o filme completo e seus trechos, como esperado.

Os resultados obtidos na avaliação de desempenho da técnica descrita neste trabalho, utilizando o computador pessoal e a métrica supracitadas, são apresentados na Tabela 4.5.

Ao analisar os resultados apresentados na Tabela 4.5, nota-se que a técnica apresenta uma grande variação de desempenho conforme o vídeo de entrada adotado. Nota-se, por exemplo, que o desempenho obtido ao segmentar o vídeo *T. Gone in Sixty Seconds* apresentou um desempenho significativamente inferior aos demais trechos de vídeo, ficando abaixo da marca de 30fps. Tal resultado deve-se à maior resolução do vídeo de entrada (1920x816), fato que ocasiona em um maior esforço do método de extração dos vetores de fluxo óptico. Além disso, o estilo de produção do filme, contendo um número elevado de transições de tomadas, tomadas com reduzida duração e uma alta taxa de movimentação, característico de filmes do gênero "ação", contribui para o pior resultado de desempenho aferido entre todos os testes realizados.

[5]http://goo.gl/YyU5e

Tabela 4.5: Resultados para o teste de desempenho ao executar a implementação da técnica com os trechos da base de vídeos.

Nome	Tempo de processamento	Taxa de processamento
T. A Beautiful Mind	1273s	41.75fps
T. Back to the Future	1235s	37.55fps
T. Gone in Sixty Seconds	3261s	25.48fps
T. Ice Age	1470s	38.39fps
T. Pirates of the Caribbean: The Curse of the Black Pearl	877s	48.10fps

Porém, mesmo com o resultado obtido pelo vídeo *T. Gone in Sixty Seconds*, nota-se que o desempenho foi superior à taxa de reprodução do vídeo. Ou seja, a velocidade de processamento variou de 106.16% a 200.41% da taxa de reprodução do vídeo. Destaca-se, ainda, que quatro dos cinco trechos analisados apresentaram desempenho superior à taxa de 30fps, utilizada, por exemplo, em alguns perfis do Sistema Brasileiro de TV Digital (SBTVD). Com isso, a técnica poderia ser utilizada para aplicações "ao vivo" do SBTVD, já que a taxa de processamento é superior, nesse caso, à taxa de reprodução do sistema em si.

Infelizmente, a maioria dos trabalhos relacionados não apresenta uma avaliação de desempenho de suas técnicas. Além disso, a indisponibilidade da implementação de cada técnica dificulta a realização de comparações em um mesmo computador. Por exemplo, Smeaton, Over e Doherty (2010) apresentam uma comparação do desempenho de técnicas de segmentação em tomadas, mas utilizando-se diferentes computadores.

4.3 Testes de Eficácia

Para os testes de eficácia, que objetiva verificar se a técnica é capaz de apresentar bons resultados para a tarefa a qual foi projetada, mediu-se a precisão, abrangência e F_1, descritos na Seção 2.4.

Usualmente, na avaliação de técnicas de segmentação em tomadas e cenas, usa-se uma determinada tolerância para indicar um verdadeiro positivo (Gu et al., 2007; Rasheed e Shah, 2005). Uma abordagem clássica é a de definir um determinado valor de tolerância em termos de número de quadros, assim, se uma transição é detectada entre o limiar de to-

lerância de uma transição da base confiável, essa transição é considerada como verdadeira positiva.

O número de quadros utilizados como tolerância, na avaliação realizada, varia conforme o tipo de segmentação. Por exemplo, a transição em tomadas necessita de um limiar de tolerância baixo, haja vista que tomadas possuem, geralmente, um número pequeno de quadros. Por outro lado, o limiar de tolerância da segmentação em cenas pode ser mais elevado, haja vista que, para o usuário, uma discrepância de alguns segundos na detecção de uma transição de cena pode ser irrelevante frente à cenas com duração de, potencialmente, diversos minutos. Nesse sentido, o limiar de tolerância é descrito nas Subseções posteriores.

Os resultados obtidos são apresentados divididos por tipo de segmentação. Os resultados da segmentação em tomadas são apresentados na Subseção 4.3.1 e os resultados para a segmentação em cenas são apresentados na Subseção 4.3.2.

4.3.1 Resultados Obtidos para a Segmentação em Tomadas

Os testes de segmentação em tomadas descrevem os resultados obtidos pela primeira fase da técnica, descrita na Seção 3.1. Com isso, ao final desta fase, obtém-se o arquivo de registro contendo as tomadas detectadas pelo algoritmo do vídeo de entrada correspondente. Com tal registro, compara-se as tomadas detectadas com a base confiável pré-definida, obtendo-se os valores de precisão, abrangência e F_1 para a transição de tomadas.

Para a segmentação em tomadas, utilizou-se como limiar de tolerância até três quadros para mais ou para menos. Ou seja, caso a transição seja detectada entre uma janela de três quadros antes e três quadros depois da transição definida como confiável, a transição é aceita como verdadeira positiva. Caso a transição seja gradual, o método se aplica ao quadro inicial e final da transição da transição de tomada.

A Tabela 4.6 apresenta os resultados obtidos ao segmentar a base de trechos de vídeos descritos na Seção 4.1 em tomadas.

Nos resultados descritos na Tabela 4.6, o valor F_1 obtido para a segmentação em tomadas foi, na média, de 92.81%. O resultado obtido no trecho do filme *A Beautiful Mind*, significativamente abaixo dos demais, pode ser explicitado pelo alto número de transições graduais como *fade in* e *fade out*, que resultaram em um maior número de falsos negativos, resultando em 82.21% de abrangência obtida pela técnica.

Por outro lado, no trecho do vídeo *Back to the Future*, por se tratar de um vídeo mais antigo e que adota uma produção mais tradicional, o número de transições graduais é baixo, além das transições abruptas serem bem demarcadas por altos contrastes entre as cores, resultando em um valor de precisão acima de 98%.

Tabela 4.6: Resultados para a segmentação em tomadas dos trechos dos vídeos utilizando três quadros de tolerância.

Nome	**Precisão**	**Abrangência**	$\boldsymbol{F_1}$
T. A Beautiful Mind	96.73%	82.21%	88.88%
T. Back to the Future	98.41%	93.93%	96.12%
T. Gone in Sixty Seconds	92.48%	93.33%	92.90%
T. Ice Age	88.41%	93.44%	90.85%
T. Pirates of the Caribbean: The Curse of the Black Pearl	98.93%	91.95%	95.31%

A Tabela 4.7 apresenta os resultados obtidos ao segmentar a base de vídeos completos descritos na Seção 4.1 em tomadas, com três quadros de tolerância.

Tabela 4.7: Resultados para a segmentação em tomada dos vídeos completos utilizando três quadros de tolerância.

Nome	**Precisão**	**Abrangência**	$\boldsymbol{F_1}$
A Beautiful Mind	96.92%	81.38%	88.47%
Back to the Future	82.59%	87.85%	85.13%
Gone in Sixty Seconds	92.43%	81.57%	86.66%
Ice Age	89.44%	92.52%	90.90%
Pirates of the Caribbean: The Curse of the Black Pearl	94.40%	87.58%	90.86%

Conforme esperado, os valores de precisão e abrangência apresentados pela técnica ao realizar a segmentação em tomadas sobre os trechos de vídeos são superiores aos resultados obtidos ao segmentar os vídeos completos, devido, principalmente, à ausência de diversas transições complexas presentes no vídeo completo, mas ausentes nos trechos selecionados. Ao considerar os filmes completos, o valor médio de F_1 obtido pela técnica foi de 88.40%, cerca de 4% menor que o obtido analisando-se apenas os trechos dos filmes (média de 92.81%).

O vídeo *A Beautiful Mind* manteve o comportamento apresentado ao se segmentar o trecho do vídeo correspondente: alta precisão e abrangência significativamente menor.

Isso se deve ao fato de que o trecho de vídeo e o vídeo inteiro apresentam uma dinâmica de transições abruptas e graduais semelhantes, com diversas transições com *fade in* e *fade out*, além de algumas transições abruptas com baixa dissimilaridade em seus histogramas. Tal característica, além disso, é responsável pelo alto número de falsos negativos, resultando na menor taxa de abrangência entre os vídeos completos avaliados.

Já o vídeo *Back to the Future*, nos resultados apresentados na Tabela 4.6, apresentou valores de precisão acima de 98% ao segmentar um trecho do referido vídeo. Porém, ao se analisar o vídeo completo, o resultado obtido para a precisão caiu para 82%. Tal redução deve-se, nesse caso, ao fato de que o trecho do vídeo representa um segmento do vídeo onde há a presença de alto contraste entre as tomadas, facilitando a segmentação em tomadas. Adicionalmente, a presença de diversos efeitos especiais no início e no final do filme, muitos dos quais contendo *flashes*, representou uma queda acentuada na precisão da técnica. Individualmente, o vídeo *Back to the Future* foi o que apresentou a maior queda no valor de F_1, porém, mantendo-se acima da taxa de 85%.

Os resultados obtidos para o vídeo *Gone in Sixty Seconds* apresentaram uma significativa redução no valor de abrangência. Embora a precisão manteve-se praticamente inalterada, a abrangência caiu ao analisarem-se segmentos de ação, com elevado número de explosões e transições, tanto abrupta como graduais, complexas. Com isso, o número de falsos negativos subiu consideravelmente, ocasionando um baixo valor de abrangência. Porém, graças ao elevado valor de precisão, o valor de F_1 manteve-se acima de 86%.

O vídeo *Ice Age*, por sua vez, apresentou uma leve variação nos valores de precisão, abrangência e F_1. A precisão apresentou um leve incremento enquanto que a abrangência apresentou uma pequena queda, ocasionando valores de F_1 semelhantes, com diferença de apenas 0.05%. Isso se deve, principalmente, ao fato de o vídeo não apresentar diferenças significativas tanto em contraste, formato de produção ou conteúdo entre o trecho selecionado e o vídeo completo.

Por fim, os resultados do vídeo *Pirates of the Caribbean: The Curse of the Black Pearl* apresentaram uma queda tanto nos valores de precisão como de abrangência, devido à maior complexidade das tomadas presentes ao final do vídeo, com cores escuras de fundo e pequena variação nos histogramas. Mesmo assim, o valor de F_1 manteve-se acima de 90%, muito próximo do melhor resultado alcançado nos vídeos completos (obtido pelo filme *Ice Age*).

Para efeitos de comparação, Smeaton, Over e Doherty (2010), apresentam um trabalho descrevendo os resultados obtidos em sete anos de atividade do TRECVid em termos de segmentação em tomadas, com ênfase ao ano de 2005. As dez melhores técnicas e o valor de F_1 obtidos são apresentados na Tabela 4.8. Deve-se destacar que a tarefa do TRECVid

voltada para a segmentação em tomadas encerrou-se no ano de 2007, o que faz com que as técnicas desenvolvidas em 2005 ainda sejam boas comparações com técnicas atuais.

Tabela 4.8: Valor F_1 para as dez melhores técnicas reportadas no TRECVid 2005 em relação à segmentação em tomadas. Adaptado de (Smeaton et al., 2010).

Grupo	$\boldsymbol{F_1}$
Tsinghua U.	89.70%
NICTA	89.20%
IBM Research	87.60%
CLIPS-IMAG	87.60%
KDDI R&D	86.50%
Marburg U.	85.90%
RMIT U.	85.30%
U. Modena	84.50%
FX Palo Alto Lab.	83.90%
City U. Hong Kong	83.10%

Nota-se que a eficácia alcançada entre as melhores técnicas descritas apresentaram baixa variação no valor F_1 alcançado. A técnica com os melhores resultados (Tsinghua U.), por exemplo, obteve apenas 0.50% a mais de F_1 que o segundo colocado (NICTA) e apenas 6.60% acima da décima técnica (City U. Hong Kong). Deve-se destacar, ainda, que a Tabela 4.8 inclui os resultados obtidos para a segmentação com transições abruptas e graduais.

Nos testes realizados, mesmo considerando o pior resultado alcançado pela técnica desenvolvida (85.13% de F_1 no filme completo *Back to the Future*), a técnica alcançou resultados comparáveis às dez melhores técnicas do TRECVid 2005. Se forem considerados os trechos dos filmes, o melhor resultado obtido supera até mesmo a melhor técnica na segmentação das transições abruptas (a técnica KDDI, com 94.20% de F_1) com vantagem de cerca de 2%.

É importante destacar que o TRECVid utilizou trechos de vídeos dos domínios *sitcom*, telejornais e documentários, diferentemente da base de vídeos do domínio filmes utilizada

neste trabalho. Destaca-se, porém, que as tomadas não apresentam grande variação entre os diferentes domínios. Por exemplo, tomadas de diálogo em um filme são similares ao âncora apresentando uma notícia em um telejornal, haja vista que a movimentação é mínima (geralmente, do rosto e da parte superior do corpo do locutor), o ambiente é estático, a iluminação é constante, entre outros. Assim, supõe-se que os resultados são similares independentemente do domínio, tornando a comparação válida.

Por fim, nota-se que a técnica com os melhores resultados no TRECVid 2005 (Tsinghua U.) utilizou técnicas como aprendizagem de máquina, histogramas de cor, detecção de *flashes* e compensação de movimento. Neste sentido, dentre as dez melhores técnicas, apenas a CLIPS-IMAG não faz uso da aprendizagem de máquina, mas, mesmo assim, realiza a segmentação utilizando a detecção de *flashs*, limiares adaptativos e compensação de movimento.

Outro trabalho, em estado-da-arte na área de segmentação em tomadas, afirma superar o melhor resultado obtido no TRECVid 2006 (Zhang e Wang, 2012), com diferença de até 4% na precisão considerando ambos os tipos de transição de tomadas (abruptas e graduais). Porém, mesmo assim, os resultados são similares aos obtidos pela técnica Tsinghua U., demonstrando que as melhorias alcançadas em cerca de 7 anos de desenvolvimento são pontuais, resultando em pequenas diferenças entre os resultados reportados.

Com isso, pode-se concluir que a técnica de segmentação em tomadas desenvolvida apresenta resultados comparáveis às técnicas em estado-da-arte, mesmo as que utilizam recursos mais avançados como a aprendizagem de máquina. Com isso, conclui-se que resultados significativos podem ser alcançados apenas com o uso de histogramas, fato que contribui para um baixo custo computacional. Por fim, destaca-se que a técnica de segmentação de tomadas apresentada neste trabalho pode ser utilizada como base para outras técnicas de segmentação em tomadas, podendo melhorar consideravelmente os resultados obtidos por tais técnicas.

4.3.2 Resultados Obtidos para a Segmentação em Cenas

Nos testes relativos à segmentação em cenas, utilizam-se as tomadas segmentadas na primeira fase do algoritmo (cujos resultados de eficácia foram apresentadas na Subseção 4.3.1), além do fluxo óptico e os histogramas de todos os quadros do vídeo.

Para os testes de segmentação em cenas, adotou-se uma tolerância equivalente a cinco segundos de vídeo. Como os vídeos da base confiável continham vídeos codificados na taxa de cerca de 24 quadros por segundo, a tolerância para a detecção de cenas adotada foi de 120 quadros para mais ou para menos. Nota-se que não foram aceitas transições múltiplas em uma mesma janela de tolerância. Por exemplo, caso uma transição de cena seja detectada entre os quadros 100 e 101 e outra transição de cenas nos quadros 150 e

151, mas a transição de base confiável seja entre os quadros 120 e 121, apenas uma das transições anteriores seria aceita como verdadeira positiva (considerando que nenhuma das duas transições detectadas possam pertencer à alguma outra janela de tolerância), sendo a outra classificada como falsa positiva.

Além disso, destaca-se que cenas com baixa relevância semântica, que não apresentem uma ação essencial ao filme, foram desconsideradas da análise. Tais cenas são caracterizadas por serem visualmente dissimilares às cenas adjacentes e servirem, geralmente, como uma forma de "transição" entre uma cena e a próxima cena, sendo que ela, isoladamente, não representa uma cena. Nas avaliações realizadas, cenas com duração menor ou igual a 5 segundos (120 quadros) foram desconsideradas da avaliação, sejam da base confiável ou dos arquivos de registro obtidos pela técnica. Tal duração foi estimada ao avaliar os trechos que apresentavam tal característica na base confiável.

Os resultados para a segmentação em cenas são apresentados e discutidos nas Subseções 4.3.2.1 (segmentação dos trechos dos filmes) e 4.3.2.2 (segmentação dos filmes completos), apresentadas a seguir.

4.3.2.1 Segmentação em Cenas dos Trechos dos Vídeos

A segmentação em cenas dos trechos de vídeo é importante, pois diversos trabalhos relacionados utilizam-se de "segmentos" pré-determinados de vídeos para realizar a avaliação de suas técnicas (Chasanis et al., 2009; Rasheed e Shah, 2003, 2005). Os resultados obtidos para a segmentação em cenas dos trechos dos vídeos, considerando a metodologia supracitada, são apresentados na Tabela 4.9.

Tabela 4.9: Resultados para a segmentação em cenas dos trechos dos vídeos da base de testes com 120 quadros de tolerância.

Nome	Precisão	Abrangência	F_1
T. A Beautiful Mind	93.10%	79.41%	85.71%
T. Back to the Future	85.00%	85.00%	85.00%
T. Gone in Sixty Seconds	81.81%	86.74%	84.21%
T. Ice Age	89.06%	89.06%	89.06%
T. Pirates of the Caribbean: The Curse of the Black Pearl	87.17%	89.47%	88.31%

Nos resultados apresentados na Tabela 4.9, o trecho do filme *T. A Beautiful Mind* apresentou resultados acima de 90% para a precisão e cerca de 80% para a abrangência.

Nesse trecho, várias cenas detectadas coincidiam com mudanças de foco entre os rostos dos personagens e o ambiente à sua volta, característica de filmes do estilo drama. Além disso, as características de cor apresentaram poucas variações de tonalidade, resultando em transições de cenas com baixa dissimilaridade em seus histogramas. Como resultado, obteve-se um baixo índice de falsos positivos, devido ao fato de as transições apresentarem alta variação nos histogramas devido a grandes diferenças de iluminação, por exemplo. Infelizmente, a abrangência sofreu o impacto do grande número de transições graduais entre as tomadas, fato também comprovado nos resultados para a segmentação em tomadas apresentada na Tabela 4.6.

No trecho do filme *T. Back to the Future* e no filme como um todo, porém, as transições de tomadas e cenas são caracterizadas por uma alta variação nos histogramas dos quadros de cada tomada. Tal característica pode ser comprovada nos resultados apresentados na Tabela 4.6, onde o trecho obteve o maior valor de F_1 dentre todos os trechos analisados. Como desvantagem, a grande variação entre os histogramas de cada tomada resultou em um maior índice de falsos positivos, ocasionando um valor de precisão significativamente inferior ao obtido na segmentação em tomadas.

O trecho do filme *T. Gone in Sixty Seconds* apresentou os piores resultados de precisão e F_1 dentre todos os trechos avaliados. No trecho selecionado, mudanças bruscas de iluminação, mudanças rápidas de foco entre diferentes cenários e cenas com alta disparidade na sua duração resultaram em resultados inferiores aos obtidos nos outros trechos. Outro fator que colabora para tais resultados é a duração consideravelmente maior do trecho frente aos demais trechos selecionados. Mesmo assim, como diversas tomadas apresentaram dinâmica de movimento semelhante, o valor de F_1 aproximou-se da marca de 85%.

O trecho da animação *T. Ice Age* apresentou resultados satisfatórios, principalmente em relação à abrangência, com valores próximos de 90%. Nesse caso, os bons resultados obtidos na segmentação em tomadas, a alta variação entre os quadros-chaves de cada tomada e o baixo índice de transições graduais colaboraram para a obtenção do melhor resultado para a segmentação em cenas.

Por fim, o trecho do filme *T. Pirates of the Caribbean: The Curse of the Black Pearl* apresentou valores satisfatórios de precisão e abrangência, além de uma baixa variação entre eles, resultando no segundo melhor índice F_1 entre todos os trechos analisados. Nesse caso, o fato de que as cenas de diálogo e de ação apresentam uma dinâmica de movimento similar, resultou em um baixo índice de falsos positivos. Além disso, a razoável variação entre as tonalidades de cada cena (resultando em histogramas significativamente dissimilares) resultou no melhor valor de abrangência dentre todos os trechos analisados.

Tais resultados, além de significativos, são semelhantes ou até mesmo superiores ao trabalho apresentado por Rasheed e Shah (2003), outro trabalho que também utilizou

a coerência visual e a análise de movimento para segmentar filmes em cenas e que se destaca por apresentar bons resultados na segmentação em cenas nesse domínio. Uma comparação dos resultados reportados pelos autores e os resultados obtidos nessa avaliação são apresentadas na Seção 4.4.

Uma importante consideração a ser realizada diz respeito ao impacto da utilização do fluxo óptico. Tal característica é utilizada para amenizar a *oversegmentation* obtida ao realizar a segmentação exclusivamente com a coerência visual, de maneira a reduzir o número de falsos positivos através da junção de cenas adjacentes com dinâmica de movimento similares e/ou acima de determinado limiar. Maiores detalhes acerca do procedimento de junção baseada no fluxo óptico são descritos na Subseção 3.2.5.

Para avaliar o impacto do fluxo óptico executou-se o algoritmo sem a presença de tal método de junção, ou seja, realizou-se a avaliação utilizando apenas a coerência visual (baseada em histogramas). Os resultados, considerando tal configuração e utilizando os trechos dos filmes, são apresentados na Tabela 4.10.

Tabela 4.10: Comparação entre os resultados obtidos com ou sem o método de junção baseado no fluxo óptico para a segmentação em cenas dos trechos dos filmes.

Filme	**Métrica**	**Sem Fluxo Óptico**	**Com Fluxo Óptico**
T. A Beautiful Mind	Precisão	87.09%	93.10%
	Abrangência	79.41%	79.41%
	F_1	83.07%	85.71%
T. Back to the Future	Precisão	75.00%	85.00%
	Abrangência	90.00%	85.00%
	F_1	81.81%	85.00%
T. Gone in Sixty Seconds	Precisão	78.94%	81.81%
	Abrangência	90.36%	86.74%
	F_1	84.26%	84.21%
T. Ice Age	Precisão	80.28%	89.06%
	Abrangência	89.06%	89.06%
	F_1	84.44%	89.06%
T. Pirates of the Caribbean: The Curse of Black Pearl	Precisão	85.00%	87.17%
	Abrangência	89.47%	89.47%
	F_1	87.17%	88.31%

A primeira consideração quanto aos resultados apresentados na Tabela 4.10 diz respeito à taxa de abrangência. Como o fluxo óptico é utilizado para unir cenas adjacentes com dinâmica de movimento similar (ou se ambas forem acima de determinado limiar), o melhor resultado esperado em termos de abrangência é a de que a taxa permaneça inalterada, ou seja, nenhuma transição verdadeira positiva seja removida. Nesse sentido,

nota-se que a abrangência permaneceu constante em três dos cinco trechos selecionados (*T. A Beautiful Mind*, *T. Ice Age* e *T. Pirates of the Caribbean: The Curse of Black Pearl*), fato que indica que o método não causou a junção de quaisquer duas ou mais cenas erroneamente nesses trechos.

Além disso, a aplicação da junção baseada na dinâmica de movimento, estimada através do fluxo óptico, resultou em um acréscimo na precisão em todos os trechos dos vídeos testados. O aumento percebido variou de 2.27% (para o vídeo *T. Pirates of the Caribbean: The Curse of Black Pearl*) a até 10.00% (para o vídeo *T. Back to the Future*). Tais diferenças podem ser explicadas devido a características particulares dos vídeos analisados. Por exemplo, o vídeo *T. Back to the Future* apresenta um grande número de cenas adjacentes com dinâmica de movimento similar, como em cenas de diálogos. Tal fato, inclusive, causa a junção errônea de algumas transições verdadeiras positivas. Percebe-se pela medida F_1 que o ganho na precisão compensa, nesse caso, a diminuição na taxa de abrangência.

No vídeo *T. Pirates of the Caribbean: The Curse of Black Pearl*, por outro lado, cenas com alta dinâmica de movimento são geralmente intercaladas por cenas com baixa dinâmica de movimento. Assim, ao analisar cenas adjacentes, o número de transições unidas é reduzido, ocasionado o ganho de apenas 2.17% na precisão. Mesmo assim, devido a essas mesmas características, nenhum verdadeiro positivo foi eliminado, mantendo a taxa de abrangência inalterada.

Em termos de F_1, apenas o vídeo *T. Gone in Sixty Seconds* apresentou uma variação negativa, mesmo que pequena, ao aplicar a junção baseada em fluxo óptico. Porém, a diferença entre os resultados é de apenas 0.05%, um valor baixo frente ao ganho de cerca de 5% obtido no vídeo *T. Ice Age*, por exemplo. Assim, pode-se concluir que a utilização da análise da dinâmica de movimento através do fluxo óptico apresentou resultados satisfatórios, de maneira a amenizar a *oversegmentation*, ao custo de uma redução na abrangência em alguns vídeos específicos.

4.3.2.2 Segmentação em Cenas dos Vídeos Completos

Além da análise de trecho pré-selecionados de vídeos, é importante realizar uma análise com os vídeos completos. Ao analisar os resultados obtidos com vídeos de maior duração e que apresentem características diferentes conforme o tempo pode-se melhor perceber as deficiências do método proposto. Assim, realizou-se os testes da segmentação em cenas dos vídeos completos, cujos resultados são apresentados na Tabela 4.11.

Os resultados apresentados na Tabela 4.11 mostram uma redução significativa nos valores de precisão, abrangência e F_1 em relação aos testes com os trechos dos filmes. Tal redução era esperada, haja vista que diversas cenas complexas não estavam incluídas

Tabela 4.11: Resultados para a segmentação em cenas dos vídeos completos da base de testes com 120 quadros de tolerância.

Nome	**Precisão**	**Abrangência**	F_1
A Beautiful Mind	59.79%	61.70%	60.73%
Back to the Future	56.84%	67.50%	61.71%
Gone in Sixty Seconds	62.87%	72.41%	67.30%
Ice Age	77.77%	86.77%	82.03%
Pirates of the Caribbean: The Curse of the Black Pearl	64.42%	81.70%	72.04%

nos trechos de vídeos, escolhidos de maneira a maximizar os resultados da técnica de segmentação em si. Destaca-se ainda que, embora todos os valores tenham sofrido uma redução, a taxa de precisão foi a que apresentou a maior redução em comparação com a avaliação utilizando os trechos dos vídeos, devido principalmente a um aumento expressivo na taxa de falsos positivos.

No filme *A Beautiful Mind*, um elevado índice de transições foram desconsideradas, resultando em uma redução significativa na abrangência. Tal fato ocorre perto do fim do filme, onde muitas das transições não são facilmente reconhecidas visualmente, resultando no baixo número de transições encontradas. Quanto à precisão, cenas como a da perseguição de carros entre os minutos 54:00 e 56:00, por exemplo, resultaram em uma detecção de diversas transições de cenas ao invés de apenas duas transições, evidenciando uma baixa robustez da técnica em cenas de ação e com conteúdo visual altamente dissimilar às tomadas anteriores.

No filme *Back to the Future*, a taxa de precisão também apresentou uma queda acentuada, resultando na pior taxa de precisão dentre todos os filmes testados, com valores próximos a 55%. Nesse caso, a introdução de cenas críticas como, por exemplo, as presentes entre os minutos 18:00 a 24:00 e entre 1:31:00 a 1:45:00 (ambos cenas escuras com a presença de efeitos especiais que modificavam radicalmente os histogramas), resultaram em uma alta taxa de falsos positivos.

É interessante notar que a segmentação em tomadas, cujos resultados são descritos na Tabela 4.7, também apresentaram uma queda acentuada na precisão do filme *Back to the Future* devido as trechos supracitados. Nesse caso, há uma correlação entre a segmentação em tomadas e em cenas haja vista que ambas são fortemente baseadas em informações obtidas através dos histogramas. Em tais segmentos, a adoção de uma técnica capaz de

reduzir o impacto dos efeitos especiais nesses trechos (*flashes*, principalmente) poderia resultar em melhoras acentuadas na precisão da técnica.

Já no filme *Gone in Sixty Seconds*, a presença de cenas de ação posteriores aos presentes no trecho do filme resultou em um decréscimo significativo no valor da precisão. Por exemplo, a cenas de perseguição entre os tempos 1:25:00 a 1:36:00 e a cena de tiroteio entre 1:41:00 a 1:45:00 gerou diversos falsos positivos, demonstrando uma deficiência da técnica para tomadas com baixo número de quadros e com grande variação entre seus histogramas. Infelizmente, mesmo com a análise da dinâmica de movimento, o número de falsos positivos manteve-se elevado.

O filme *Ice Age*, por sua vez, apresentou a melhor taxa de abrangência entre todos os filmes avaliados, alcançando valores acima de 85%, contribuindo para o melhor valor de F_1 dentre todos os filmes completos analisados. Tais resultados são possíveis devido ao fato de a animação apresentar uma produção constante com cenas de ação, diálogos e transições bem demarcadas com mudanças significativas na informação de cor e iluminação, resultando em histogramas dissimilares. Assim, grande parte das transições de cenas na animação pôde ser corretamente detectada. Isso também contribuiu para a obtenção do melhor valor para a precisão, acima de 75%. Infelizmente, graças à presença de algumas cenas de ação com mudanças rápidas de tomadas, como por exemplo, entre os trechos 14:00 a 17:00, 27:00 a 31:00, 45:00 a 48:00 e 1:02:00 a 1:06:00, o algoritmo apresentou um número considerável de falsos positivos.

Por fim, o filme *Pirates of the Caribbean: The Curse of the Black Pearl* apresentou o segundo melhor resultado dentre todos os testes com os filmes completos, alcançando mais de 80% de abrangência. Assim como os demais filmes, a técnica mostrou-se deficiente em trechos onde há variações bruscas entre os histogramas dos quadros-chaves aliadas a tomadas de baixa duração, causando um aumento excessivo no número de falsos positivos.

Com os testes realizados sobre a base de vídeos, do domínio de filmes, pode-se concluir que a técnica apresenta resultados satisfatórios na segmentação em tomadas, alcançando, no pior caso, 85% de F_1 no filme completo *Back to the Future*, e, no melhor caso, alcançando a marca de 96% de F_1, desta vez no trecho do filme *T. Back to the Future*. Para a segmentação em cenas de trechos selecionados, o resultado obtido é inferior ao alcançado na segmentação em tomadas, mas, mesmo assim, ultrapassando 84% de F_1 no pior caso (89% no melhor caso).

Porém, caso a segmentação de cenas seja realizada em filmes completos, sem qualquer seleção prévia do conteúdo a ser analisado, o resultado apresenta uma queda bastante acentuada nas taxas de precisão, abrangência e F_1. Nesse caso, o vídeo *Back to the Future*, por exemplo, atinge valores próximos a 60% de precisão, abrangência e F_1, o pior

valor individual dentre todos os testes realizados. Valores baixos de precisão indicam um elevado número de falsos positivos, caracterizando a *oversegmentation*.

Mesmo assim, os resultados de abrangência são consideravelmente melhores que os obtidos em termos de precisão. Isso indica que a maioria das transições de cenas pode ser encontrada com a técnica, mesmo que para isso transições inexistentes também sejam detectadas. Tal comportamento abre a possibilidade de novas pesquisas, haja vista que procedimentos posteriores à técnica poderiam reduzir a taxa de falsos positivos, causando um aumento na taxa de precisão e, consequentemente, no valor F_1.

Por fim, assim como apresentado na Subseção 4.3.2.1, é interessante uma análise do impacto da utilização ou não do fluxo óptico, a característica utilizada para o procedimento de junção de cenas adjacentes com dinâmica de movimento similares. A Tabela 4.12 descreve os resultados com e sem a utilização do fluxo óptico para tal junção.

Tabela 4.12: Comparação entre os resultados obtidos com ou sem o método de junção baseado no fluxo óptico para a segmentação em cenas dos vídeos completos.

Filme	**Métrica**	**Sem Fluxo Óptico**	**Com Fluxo Óptico**
A Beautiful Mind	Precisão	58.65%	59.79%
	Abrangência	64.89%	61.70%
	F_1	61.61%	60.73%
Back to the Future	Precisão	51.92%	56.84%
	Abrangência	67.50%	67.50%
	F_1	58.69%	61.71%
Gone in Sixty Seconds	Precisão	61.66%	62.87%
	Abrangência	76.55%	72.41%
	F_1	68.30%	67.30%
Ice Age	Precisão	72.41%	77.77%
	Abrangência	86.77%	86.77%
	F_1	78.94%	82.03%
Pirates of the Caribbean: The Curse of Black Pearl	Precisão	60.88%	64.42%
	Abrangência	83.53%	81.70%
	F_1	70.43%	72.04%

Como explicitado anteriormente, a aplicação do procedimento de junção baseada na dinâmica de movimento tende a produzir uma queda na abrangência, haja vista a remoção de transições de cenas realizada pelo procedimento. Assim como nos resultados ao segmentar os trechos dos filmes, os vídeos *Back to the Future* e *Ice Age* não apresentaram queda nas taxas de abrangência, mesmo após a aplicação da análise da dinâmica de movimento baseada no fluxo óptico. Isso, aliado ao aumento de cerca de 5% na precisão resultou em um melhora significativa no valor de F_1 contribuindo, por exemplo, para

que o filme *Ice Age* obtivesse o melhor resultado de F_1 dentre todos os filmes completos testados.

Infelizmente, ao aplicar a junção baseada na dinâmica de movimento, houve uma redução perceptível da taxa de F_1 nos filmes *A Beautiful Mind* e *Gone in Sixty Seconds*. Ao contrário da queda de 0.05% percebida ao segmentar o trecho do vídeo *Gone in Sixty Seconds*, a redução ao segmentar os vídeos completos foi de cerca de 1%, um valor mais significativo. Tal resultado no filme *Gone in Sixty Seconds*, assim como na segmentação do trecho do filme *T. Gone in Sixty Seconds*, pode ser explicado pela grande variação entre cenas de alta movimentação, seguidas por cenas com baixa movimentação.

O filme *A Beautiful Mind*, por sua vez, apresenta baixa variação na dinâmica de movimento de cenas adjacentes, resultando na junção errônea de diversas cenas, reduzindo significativamente o número de verdadeiros positivos. Isso indica que o procedimento de junção baseada na dinâmica de movimento poderia ser aperfeiçoado de maneira a considerar outros fatores além da dinâmica de movimento, como a própria coerência visual, de maneira a evitar os efeitos colaterais da aplicação da junção baseada no fluxo óptico.

4.4 Comparação da Segmentação em Cenas com Trabalhos Relacionados

Além de apresentar os resultados da técnica desenvolvida, é importante também compará-la com outras técnicas em estado-da-arte. Para tal, selecionou-se duas técnicas de segmentação em cenas no estado-da-arte, a técnica BSC desenvolvida por Rasheed e Shah (2003) e a técnica desenvolvida por Chasanis et al. (2009).

A técnica BSC utiliza-se da coerência visual e de vetores de movimento de tomadas adjacentes para realizar a segmentação em cenas. Na BSC, os procedimentos realizados para obter a segmentação em cenas são similares aos procedimentos adotados na técnica desenvolvida, que se baseia em segmentar o vídeo em tomadas, selecionar um conjunto de quadros-chave, calcular a coerência visual entre as tomadas e reduzir o número de falsos positivos através de métodos heurísticos baseados na coerência visual e/ou na dinâmica de movimento. Porém, a técnica desenvolvida apresenta diversas diferenças em relação a técnica BSC tais como a adoção do fluxo óptico para o cálculo da dinâmica de movimento (ao invés dos vetores de movimento), um método de segmentação em tomadas mais rebuscado, cálculo diferenciado da coerência visual, entre outros.

Deve-se destacar que a técnica BSC, mesmo tendo sida apresentada em 2003, ainda é considerada em estado-da-arte pois mesmo trabalhos relacionados mais recentes, como o desenvolvido por Chen e Li (2010), obtém resultados inferiores se comparados com a técnica BSC.

A técnica desenvolvida por Chasanis et al. (2009), doravante chamado de *Bag of Visual Words* (BVW), por outro lado, baseia-se na extração de características locais SIFT (Lowe, 1999) e CCH (Huang et al., 2006, 2008) dos quadros-chave de cada tomada, seguida da criação de dicionários visuais para a segmentação em cenas. Depois de criado o dicionário visual, para cada tomada é calculado um histograma suavizado de "palavras visuais", sendo que as cenas são detectadas localizando-se máximos locais dos valores da distância euclidiana entre tomadas adjacentes.

Outro motivo para terem sido selecionadas, além de serem técnicas em estado-da-arte, é a de que tais técnicas utilizaram filmes em comum para a avaliação da precisão, abrangência e F_1 no domínio filmes. A técnica BSC realizou avaliações sobre trechos dos filmes *Terminator 2*, *Golden Eye*, *Gone in Sixty Seconds*, *Top Gun* e *A Beautiful Mind*. Os autores da técnica BVW, por sua vez, realizaram as avaliações com trechos dos filmes *A Beautiful Mind*, *Sex and the City* e *Gone in Sixty Seconds*. Deve-se destacar que os trechos selecionados em cada filme possuem durações diferentes (para o trecho do filme *Gone in Sixty Seconds*) ou número de transições de tomada diferentes (para o trecho do filme *A Beautiful Mind*), o que indica que não foram utilizados os mesmos trechos.

Como as técnicas avaliaram a eficácia da segmentação em cenas utilizando trechos dos filmes, comparou-se tais valores com os resultados apresentados na Tabela 4.9. Para isso, supõe-se que os autores selecionaram os trechos dos filmes que apresentam os melhores resultados para cada técnica, de maneira análoga aos trechos selecionados neste trabalho.

Na Tabela 4.13 é apresentada uma comparação entre a técnica BSC desenvolvida por Rasheed e Shah (2003) e os resultados alcançados na técnica desenvolvida.

Tabela 4.13: Comparação entre a técnica BSC desenvolvida por Rasheed e Shah (2010) e a técnica desenvolvida

Filme	**Métrica**	**BSC**	**Técnica Desenvolvida**
T. A Beautiful Mind	Precisão	71.40%	93.10%
	Abrangência	88.20%	79.41%
	F_1	78.91%	85.71%
T. Gone in Sixty Seconds	Precisão	76.70%	81.81%
	Abrangência	84.60%	86.74%
	F_1	80.45%	84.21%

Os resultados apresentados na Tabela 4.13 mostram um equilíbrio entre os valores alcançados pela técnica BSC e pela técnica desenvolvida. No trecho do filme *T. A Beautiful Mind*, por exemplo, a técnica desenvolvida apresentou um incremento de mais de 20% de precisão em relação a técnica BSC obtendo, porém, uma abrangência 8% menor. Assim, a técnica desenvolvida obteve melhores resultados segundo a métrica F_1 (85.71% da

técnica desenvolvida contra 78.91% da técnica BSC). Para o trecho do filme *T. Gone in Sixty Seconds*, por outro lado, a técnica BSC apresentou resultados inferiores na precisão, abrangência e F_1.

Deve-se destacar ainda que o trecho do filmes *T. Gone in Sixty Seconds* é o que obteve o pior resultado de F_1 na técnica desenvolvida (como apresentado na Tabela 4.9). Na técnica BSC, por exemplo, o melhor resultado foi alcançado em um trecho do filme *Terminator 2*, obtendo valor de F_1 de 83.78%, abaixo dos resultados obtidos pela técnica desenvolvida em qualquer dos trechos dos vídeos avaliados. Além disso, o pior resultado alcançado pela técnica BSC (73.36% de F_1 em um trecho do filme *Golden Eye*) é bastante inferior ao pior resultado alcançado pela técnica desenvolvida (84.21% em *T. Gone in Sixty Seconds*).

Por fim, cita-se que a técnica desenvolvida não é restrita a vídeos de determinados formatos. No caso da BSC, utilizam-se os vetores de movimento extraídos do fluxo de entrada MPEG-1 para o cálculo de dinâmica de movimento, limitando a técnica a vídeos de tal formato. Na técnica desenvolvida, porém, tal análise é realizada utilizando o fluxo óptico que é extraído diretamente do vídeo decodificado. Assim, a técnica desenvolvida não apresenta a limitação imposta pela técnica BSC e, ainda assim, obtém resultados semelhantes.

Os autores da técnica BVW (Chasanis et al., 2009) realizaram avaliações utilizando diferentes tamanhos de dicionários SIFT e CCH (10, 20, 50, 100, 200 ou 500 palavras visuais). Para efeitos de comparação, selecionou-se os melhores resultados individuais para cada vídeo, que são os resultados obtidos utilizando SIFT com 100 palavras visuais (para o vídeo *A Beautiful Mind*) e o SIFT com 200 palavras visuais (para o vídeo *Gone in Sixty Seconds*). A comparação entre a técnica BVW e a técnica desenvolvida é apresentada na Tabela 4.14.

Tabela 4.14: Comparação entre a técnica BVW desenvolvida por Chasanis et al. (2009) e a técnica desenvolvida

Filme	**Métrica**	**BVW**	**Técnica Desenvolvida**
T. A Beautiful Mind	Precisão	88.89%	93.10%
	Abrangência	88.89%	79.41%
	F_1	88.89%	85.71%
T. Gone in Sixty Seconds	Precisão	73.03%	81.81%
	Abrangência	87.84%	86.74%
	F_1	79.75%	84.21%

Os resultados mostram que a técnica BVW obteve resultados similares aos obtido na técnica desenvolvida. No trecho do filme *T. A Beautiful Mind*, por exemplo, a técnica

desenvolvida superou a técnica de Chasanis et al. (2009) em termos de precisão (com diferença acima de 4.0%), perdendo em termos de abrangência e F_1. No trecho do filme *T. Gone in Sixty Seconds* a técnica desenvolvida obteve melhores taxas de precisão e F_1 quando comparada a técnica desenvolvida Chasanis et al. (2009), embora a técnica BVW tenha obtido uma melhor taxa de abrangência.

Os resultados demonstrados na Tabela 4.14 mostram que a técnica desenvolvida é comparável a técnica BVW, mantendo como uma de suas principais características o baixo custo computacional. Portanto, pode-se concluir que a técnica desenvolvida é capaz de alcançar resultados significativos na segmentação em cenas de vídeos do domínio filmes sem necessitar de informações particulares do formato de vídeo de entrada, como a técnica BSC desenvolvida por Rasheed e Shah (2003) e, também, sem precisar recorrer à técnicas complexas e com alto custo computacional, como a técnica BVW desenvolvida por Chasanis et al. (2009).

4.5 Limites da Técnica Desenvolvida

Embora os resultados apresentados na Seção 4.4 mostram que a técnica desenvolvida supera a técnica baseada na coerência visual em estado-da-arte, além de obter resultados similares à técnica BVW, é importante destacar as limitações que a técnica desenvolvida apresenta. Nessa seção serão destacadas as principais limitações encontradas para a segmentação em cenas, objetivo da técnica desenvolvida.

Os resultados apresentados na Subseção 4.3.2 demonstram que, em vídeos específicos, a técnica obtém resultados expressivos para a segmentação em cenas, alcançando cerca de 85% de F_1. Porém, considerando os filmes completos, a eficácia cai substancialmente, chegando ao patamar de 60% de F_1 (mas mantendo 82% no filme *Ice Age*). Em tais casos, o aumento da complexidade das cenas ocasiona diversos erros como falsos positivos e falsos negativos devido a tomadas adjacentes substancialmente dissimilares, cenas com duração variável e outros fatores.

Nesse sentido, a técnica desenvolvida estima as transições de cenas considerando a coerência baseada em histogramas dos quadros-chave extraídos de cada tomada. Caso duas tomadas adjacentes sejam similares (de maneira mais precisa: pelo menos um par de quadros-chave apresentar alto valor de intersecção de histogramas), elas provavelmente serão consideradas como pertencentes à uma única cena. Se tal afirmação não for verdadeira (ou seja, as tomadas são similares mas pertencem a cenas diferentes), a técnica será incapaz de detectar tal transição de cenas: todos os procedimentos posteriores visam fundir cenas adjacentes, e não criar novas transições de cenas.

Outra limitação diz respeito ao caso onde um segmento dissimilar é detectado entre uma mesma cena, gerando falsos positivos. Embora os procedimentos de junção de cenas descritos no Capítulo 3 procurem resolver tal problema, a maioria deles opera através da análise de histogramas. Caso o segmento seja bastante dissimilar aos segmentos adjacentes (indicado por um baixo valor de BSCW para o segmento), tais procedimentos serão incapazes de corrigir tal problema. A única forma de eliminar tais falsos positivos é através dos últimos dois procedimentos desenvolvidos: a junção de cenas com dinâmica de movimento similar e a junção de cenas intercaladas por cenas dissimilares. Porém, ambos os procedimentos são conservadores: os limiares utilizados em ambos os procedimentos foram estimados de forma a evitar a junção de transições verdadeiras positivas. Por exemplo, as Tabelas 4.12 e 4.10 mostram que a junção baseada no fluxo óptico tende a conservar a taxa de abrangência, ou seja, nesses casos, nenhum verdadeiro positivo foi eliminado.

Embora tal cenário seja hipotético, na prática, diversas cenas no domínio de filmes apresentam tal característica. Os trechos selecionados e avaliados apresentam como característica principal o fato de as cenas serem bem delimitadas, ou seja, entre uma mesma cena todas as tomadas apresentam alta similaridade visual. Os resultados obtidos com tais trechos, como descritos na Tabela 4.9, são bastante superiores ao obtido ao segmentar os filmes completos.

De maneira abstrata, tais erros ocorrem devido à lacuna semântica entre o conceito de "cena" e as características de baixo nível extraídas do vídeo. Embora as cores e a dinâmica de movimento sejam bons indicativos para determinar a transição de cenas, tais características não determinam a transição de cenas. Ou seja, um mesmo segmento de um filme pode conter trechos dissimilares entre si (em termos de cores e/ou de movimento) e ainda sim pertencerem a uma mesma cena. Sendo o contrário também válido: segmentos visualmente similares (em termos de cores e de dinâmica de movimento) podem pertencer a cenas distintas. Ou seja, em um domínio complexo tal como o de filmes, basear-se apenas em características de histogramas e fluxo óptico pode ser insuficiente para determinar corretamente a transição de cenas.

4.6 Considerações finais

Este capítulo apresentou uma avaliação realizada sobre a técnica descrita no Capítulo 3 em termos de desempenho, o quão rápido uma implementação do algoritmo pode executar, e de eficácia, o quão bom é o resultado obtido pela técnica. Para isso, descreveu-se na Seção 4.1 a base confiável adotada para os testes realizados. A Seção 4.2 apresentou os resultados de desempenho de uma implementação da técnica, as Subseções 4.3.1 e 4.3.2 apresentaram

os resultados de eficácia da segmentação em tomadas e em cenas, respectivamente. Por fim, na Seção 4.4 comparou-se a técnica desenvolvida com outras técnicas de segmentação em cenas em estado-da-arte e a Seção 4.5 descreveu algumas das principais limitações encontradas da técnica desenvolvida em relação à segmentação em cenas.

Em relação à base de vídeos utilizada, é importante destacar as seguintes considerações:

- O principal objetivo de se utilizar os trechos dos filmes é o de apresentar os resultados obtidos da técnica no melhor caso. Ou seja, escolheram-se trechos com características que facilitam a detecção correta de tomadas e cenas como segmentos com alto contraste, baixo número de transições graduais e alta similaridade entre tomadas de mesma cena, entre outros.

- Destaca-se que, embora os vídeos apresentem diferentes resoluções, a técnica não é dependente da resolução utilizada. De fato, apenas o número de vetores de fluxo óptico é influenciado pela resolução do vídeo, que é calculado automaticamente conforme descrito na Subseção 3.2.5. Os resultados obtidos mostram, porém, que a adoção de fluxos de vídeo com maior resolução impactam negativamente o desempenho da implementação da técnica, devido ao maior número de *pixels* a serem analisados pelo procedimento de cálculo do fluxo óptico.

- Embora seja utilizado poucos vídeos, destaca-se que diversos trabalhos relacionados também realizam os testes sobre um conjunto restrito de vídeos de entrada. Rasheed e Shah (2003), por exemplo, utilizam-se de cinco trechos de filmes e um *sitcom* com duração de 21 minutos. Chasanis et al. (2009), por sua vez, utilizaram-se de apenas três trechos de filmes para realizar sua avaliação de precisão e abrangência.

- A utilização dos filmes completos (com a remoção dos créditos de início e fim do filme) visa a apresentar uma avaliação mais criteriosa da que apresentada por trabalhos similares, que normalmente não especificam claramente qual o trecho utilizado para a análise, dificultando a reprodução dos resultados obtidos.

Quanto aos resultados de desempenho apresentados, destaca-se que a técnica foi capaz de superar a taxa de reprodução (24 quadros por segundo) em todos os vídeos analisados. O pior resultado obtido, no trecho do filme *T. Gone in Sixty Seconds* é devido à maior resolução adotada no fluxo de vídeo, como explicado anteriormente.

Quanto a eficácia da técnica em relação à segmentação em tomadas, a técnica desenvolvida apresentou resultados idênticos a uma técnica em estado-da-arte (Zhang e Wang, 2012) e outras técnicas mais tradicionais (Smeaton et al., 2010). Embora o domínio utilizado seja diferente, supõe-se que as tomadas em si não apresentem uma grande variação entre os diferentes domínios, tornando possível então a comparação entre as técnicas.

Nesse sentido, a segmentação em tomadas da técnica desenvolvida alcança valores de F_1 (de 88.47% a 90.90%) similares à da técnica desenvolvida por Zhang e Wang (2012), que obteve 88.3% de F_1 considerando transições abruptas e graduais. Um resultado expressivo, haja vista a simplicidade do procedimento de segmentação em tomadas desenvolvido, baseado em histogramas, uma características bem conhecida e barata computacionalmente.

Quanto a segmentação em cenas, destaca-se que, em comparação com técnicas de segmentação de vídeo que se utilizam apenas de informações visuais, como a BSC (Rasheed e Shah, 2003) e a BVW(Chasanis et al., 2009), a técnica apresentou resultados similares, ou até mesmo superiores à ambas as técnicas. A BSC, por exemplo, obteve taxas de F_1 abaixo das obtidas no trabalho desenvolvido nos trechos dos mesmos filmes. Já a técnica desenvolvida por Chasanis et al. (2009) superou em 3% a técnica desenvolvida no trecho do filme *T. A Beautiful Mind*, mas obteve cerca de 5% a menos de F_1 no trecho do filme *T. Gone in Sixty Seconds*. Assim, a técnica desenvolvida obteve resultados superiores aos reportados por técnicas anteriores que se utilizaram de coerência visual e resultados similares aos obtidos por técnicas mais complexas, como a *Bag of Features*.

A análise dos filmes completos mostrou a presença de deficiências da técnica desenvolvida para a segmentação em cenas. A principal limitação, como apresentada na Seção 4.5, é a alta dependência de informações de cores (histogramas) e de movimento da técnica, que podem não ser bons indicativos para a segmentação em cenas de vídeos complexos, como os do domínio filmes.

Capítulo

5

Conclusões e Trabalhos Futuros

5.1 Considerações Iniciais

O objetivo deste trabalho foi o desenvolvimento de uma técnica de segmentação de vídeo em cenas, baseada em coerência visual, que seja independente do formato de vídeo de entrada.

Para isso, extraíram-se duas características de vídeo, os histogramas e o fluxo óptico, para estimar a coerência visual e a dinâmica de movimento, respectivamente. O histograma é uma característica simples que descreve a distribuição das cores em uma dada imagem, sendo a mesma adotada em diversos procedimentos da técnica, como a segmentação em tomadas, seleção dos quadros-chave, cálculo da coerência visual, entre outros. Já o fluxo óptico, obtido através da análise de quadros adjacentes, foi utilizado com o objetivo de inserir uma análise de dinâmica de movimento na técnica, haja vista a presença de erros de segmentação em cenas com alta movimentação.

As próximas seções deste capítulo descrevem os resultados obtidos deste trabalho (Seção 5.2), além de algumas sugestões para trabalhos futuros (Seção 5.3).

5.2 Resultados Obtidos

A técnica desenvolvida, em contraste com trabalhos relacionados, utilizou-se apenas de histogramas para a segmentação em tomadas, o que permite um processo de segmentação

com boa velocidade e baixo custo computacional. Assim, desenvolveu-se um método de segmentação em tomadas mais rebuscado, com a presença de duas medidas de dissimilaridade, a intersecção e a distância euclidiana entre histogramas, duas janelas deslizantes de tamanho variável com limiares independentes e calculados automaticamente. Os resultados alcançados, como descritos no Capítulo 4, são similares à outras técnicas de segmentação em tomadas no estado-da-arte, sem a necessidade de utilizar características ou algoritmos complexos e computacionalmente caros.

Em relação á segmentação em cenas, os resultados obtidos mostram que a técnica desenvolvida é superior à técnica de segmentação em cenas baseada em coerência visual em estado-da-arte, além de apresentar resultados comparáveis a outra técnica baseada em *Bag of Features*. Analisou-se ainda a segmentação dos filmes completos, evidenciando as limitações da técnica, haja vista a acentuada queda nos valores de precisão, abrangência e F_1, medidas amplamente utilizadas em trabalhos relacionados e adotadas para avaliar a técnica desenvolvida. Em relação às limitações da técnica, a Seção 4.5 apresenta algumas justificativas para tal comportamento, como a presença de cenas complexas com a inserção de tomadas altamente dissimilares.

As principais contribuições deste trabalho são:

- Uma técnica de segmentação automática em tomadas e cenas com baixo custo computacional e independência de informações extras, como formato de vídeo de entrada e domínio do vídeo. Algumas contribuições específicas da técnica são apresentadas a seguir:

    1. O desenvolvimento de uma técnica de seleção de quadros-chave capaz de selecionar um subconjunto de quadros dissimilares, mas que apresentam informações complementares. Com isso, espera-se que o conjunto de quadros-chave represente adequadamente diferentes tipos de tomadas, com enfoque especial àquelas que apresentarem grande mudança de cenário e/ou movimentação da câmera e dos objetos.

    2. Introdução de um fator no cálculo da coerência visual, chamado de memória temporal, com o objetivo de priorizar determinadas tomadas de acordo com sua posição relativa. Assim, cenas de menor duração (com poucas tomadas) podem ser detectadas mais facilmente do que em técnicas relacionadas, causando um aumento significativo na taxa de abrangência.

    3. Desenvolvimento de técnicas heurísticas, para o domínio de filmes, capazes de remover erros comuns na segmentação de cenas baseada em coerência visual, como a detecção de cenas com apenas uma tomadas, por exemplo.

4. Um método capaz de estimar a "quantidade de movimento" dentro de cada cena, utilizando o fluxo óptico, de maneira a substituir os vetores de movimento que causam a limitação de técnicas relacionadas a formatos específicos de vídeo de entrada (Chen e Li, 2010; Rasheed e Shah, 2003).

5. Cálculo automático de diversos parâmetros necessários para a segmentação em tomadas e em cenas, tornando a técnica não paramétrica.

- A implementação da técnica torna-se uma importante ferramenta para outras aplicações na área de Personalização e Adaptação de Conteúdo, haja vista que pode ser facilmente incorporada em tais aplicações para prover a segmentação em tomadas e/ou cenas de um vídeo de entrada.

- A pesquisa realizada e a técnica desenvolvida resultam no aprofundamento dos conhecimentos do grupo de pesquisa em relação a métodos de segmentação de vídeo.

- A técnica desenvolvida resultou em duas publicações, um artigo curto e um artigo completo, publicadas no Simpósio Brasileiro de Multimedia e Web (WebMedia) nos anos 2012 e 2013, respectivamente.

Além disso, a criação da base de vídeos, formada pelos vídeos e também pela *ground truth*, é importante já que possibilita que outros trabalhos da área possam utilizá-la como uma *baseline*, facilitando a comparação entre diferentes técnicas de segmentação em tomadas e/ou em cenas.

Por fim, destaca-se que a técnica desenvolvida apresentou um bom desempenho, sendo capaz de processar filmes em tempo real, excedendo 30 quadros por segundo em filmes em resolução HD (*High Definition*). Com isso, facilita-se o desenvolvimento de aplicações da Personalização e Adaptação de Conteúdo voltados para os usuários, como por exemplo, mecanismos de busca e seleção de cenas específicas.

5.3 Trabalhos Futuros

Diversos trabalhos futuros podem ser desenvolvidos de maneira a continuar a pesquisa realizada. Assim, esta seção descreve alguns possíveis trabalhos futuros que foram imaginados durante o desenvolvimento deste trabalho.

Embora a técnica tenha sido desenvolvida de forma a ser independente do domínio de vídeos de entrada, é importante realizar avaliações em outros domínios. Com isso, espera-se que mudanças significativas devam ser realizadas nos métodos heurísticos, realizados após o cálculo das cenas através da coerência visual.

Um importante trabalho futuro diz respeito ao procedimento de junção de cenas baseada na dinâmica de movimento. A análise de fluxo óptico, como desenvolvido, desconsidera qualquer outra informação presente em cenas adjacentes. Assim, por exemplo, cenas visualmente diferentes podem ser unidas indiscriminadamente, resultando em baixa abrangência e resultados inadequados para possíveis usuários. Embora não se tenha percebido tal comportamento com a base de vídeos utilizada, em outros domínios (como o de telejornais), tal comportamento deve ser levado em consideração.

O desempenho da técnica foi bastante influenciado pelo método de cálculo dos vetores de fluxo óptico. Uma possível melhoria nesse sentido, sem a necessidade de maiores alterações no método, é o de realizar o cálculo sobre as imagens com resolução reduzida. Com isso, espera-se que o desempenho da técnica melhore consideravelmente, visto a redução no número de *pixels* a serem analisados, sem comprometer os resultados devido à característica do fluxo óptico ser independente da resolução.

O último procedimento da técnica, o método de junção de cenas similares intercaladas por cenas dissimilares, poderia ser aprimorado com a introdução de uma análise com janelas deslizantes de tamanho variável. Com isso, seria possível a detecção de cenas mais complexas, como as intercaladas por mais de uma cena dissimilar.

Diversos parâmetros foram estimados tendo como base o domínio de filmes. Por exemplo, no procedimento de junção de cenas com dinâmica de movimento similar, provavelmente cenas de domínios onde a movimentação é controlada, como em telejornais, as cenas jamais serão classificadas como de "alta movimentação". Tal comportamento deve-se ao procedimento utilizar-se da resolução do vídeo de entrada para estimar se um determinado segmento possui ou não alta movimentação. Assim, uma maior análise desse e de outros parâmetros pode ser realizada com o objetivo de encontrar limiares adequados a mais de um domínio de vídeo.

Referências

Adomavicius, G.; Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, v. 17, n. 6, p. 734–749, 2005. Disponível em `http://dx.doi.org/10.1109/TKDE.2005.99` (Acessado em 09/01/2014)

Ahmed, N.; Natarajan, T.; Rao, K. Discrete cosine transform. *IEEE Transactions on Computers*, v. C-23, n. 1, p. 90–93, 1974. Disponível em `http://dx.doi.org/10.1109/T-C.1974.223784` (Acessado em 09/01/2014)

Aigrain, P.; Joly, P. The automatic real-time analysis of film editing and transition effects and its applications. *Computers & Graphics*, v. 18, n. 1, p. 93–103, 1994. Disponível em `http://dx.doi.org/10.1016/0097-8493(94)90120-1` (Acessado em 09/01/2014)

Altheide, D. L. *Media power*. Sage Publications, Beverly Hills, 288 p., 1985.

Ariki, Y.; Kumano, M.; Tsukada, K. Highlight scene extraction in real time from baseball live video. In: *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval*, New York, NY, USA: ACM, 2003, p. 209–214. Disponível em `http://doi.acm.org/10.1145/973264.973297` (Acessado em 09/01/2014)

Baber, J.; Satoh, S.; Afzulpurkar, N.; Keatmanee, C. Bag of visual words model for videos segmentation into scenes. In: *Proceedings of the 5th International Conference on Internet Multimedia Computing and Service*, New York, NY, USA: ACM, 2013, p. 191–194.

Disponível em http://doi.acm.org/10.1145/2499788.2499814 (Acessado em 09/01/2014)

Baeza-Yates, R.; Ribeiro-Neto, B. *Modern information retrieval.* Addison Wesley, 544 p., 1999.

Bai, L.; Lao, S.-Y.; Liu, H.-T.; Bu, J. Video shot boundary detection using petri-net. In: *International Conference on Machine Learning and Cybernetics*, 2008, p. 3047–3051.
Disponível em http://dx.doi.org/10.1109/ICMLC.2008.4620930 (Acessado em 09/01/2014)

Barron, J.; Fleet, D.; Beauchemin, S.; Burkitt, T. A. Performance of optical flow techniques. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1992, p. 236–242.
Disponível em http://dx.doi.org/10.1109/CVPR.1992.223269 (Acessado em 09/01/2014)

Besiris, D.; Laskaris, N.; Fotopoulou, F.; Economou, G. Key frame extraction in video sequences: a vantage points approach. In: *9th IEEE Workshop on Multimedia Signal Processing*, MMSP '07, 2007, p. 434–437.
Disponível em http://dx.doi.org/10.1109/MMSP.2007.4412909 (Acessado em 09/01/2014)

Bouguet, J.-y. Pyramidal implementation of the lucas kanade feature tracker. *Intel Corporation, Microprocessor Research Labs*, 2000.
Disponível em http://robots.stanford.edu/cs223b04/algo_tracking.pdf (Acessado em 09/01/2014)

Brunelli, R.; Mich, O.; Modena, C. A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, v. 10, n. 2, p. 78–112, 1999.
Disponível em http://dx.doi.org/10.1006/jvci.1997.0404 (Acessado em 09/01/2014)

CAIS Burnt city's ancient animation on display. The Circle of Ancient Iranian Studies, 2007.
Disponível em http://www.cais-soas.com/News/2007/May2007/28-05.htm (Acessado em 09/01/2014)

Cattelan, R. G.; Teixeira, C.; Goularte, R.; Pimentel, M. D. G. C. Watch-and-comment as a paradigm toward ubiquitous interactive video editing. *ACM Transactions on*

Multimedia Computing, Communications, and Applications, v. 4, n. 4, p. 28:1–28:24, 2008.
Disponível em `http://doi.acm.org/10.1145/1412196.1412201` (Acessado em 09/01/2014)

Chaisorn, L.; Chua, T.-S.; Lee, C.-H. A multi-modal approach to story segmentation for news video. *World Wide Web*, v. 6, n. 2, p. 187–208, 2003.
Disponível em `http://dx.doi.org/10.1023/A:1023622605600` (Acessado em 09/01/2014)

Chapman, N. P.; Chapman, J. *Digital multimedia*. 3 ed. New York, NY, USA: John Wiley & Sons, Inc., 732 p., 2009.

Chasanis, V.; Kalogeratos, A.; Likas, A. Movie segmentation into scenes and chapters using locally weighted bag of visual words. In: *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, New York, NY, USA: ACM, 2009, p. 35:1–35:7.
Disponível em `http://doi.acm.org/10.1145/1646396.1646439` (Acessado em 09/01/2014)

Chasanis, V.; Likas, A.; Galatsanos, N. Efficient video shot summarization using an enhanced spectral clustering approach. In: *Proceedings of the 18th International Conference on Artificial Neural Networks*, Berlin: Springer Berlin Heidelberg, 2008, p. 847–856.
Disponível em `http://dx.doi.org/10.1007/978-3-540-87536-9_87` (Acessado em 09/01/2014)

Chen, H.; Li, C. A practical method for video scene segmentation. In: *3rd IEEE International Conference on Computer Science and Information Technology*, 2010, p. 153–156.
Disponível em `http://dx.doi.org/10.1109/ICCSIT.2010.5564930` (Acessado em 09/01/2014)

Chen, W.; Yao, L.; Zhou, J.; Dong, H. A fast geometry figure recognition algorithm based on edge pixel point eigenvalues. In: *Proceedings of the 3rd International Symposium on Computer Science and Computational Technology*, 2010, p. 297–300.

Chu, W.-T.; Lin, C.-C.; Yu, J.-Y. Using cross-media correlation for scene detection in travel videos. In: *Proceedings of the ACM International Conference on Image and Video Retrieval*, New York, NY, USA: ACM, 2009, p. 18:1–18:7.

Disponível em http://doi.acm.org/10.1145/1646396.1646420 (Acessado em 09/01/2014)

Coimbra, D. *Segmentação de cenas em telejornais: uma abordagem multimodal.* Dissertação (mestrado em ciência da computação e matemática computacional), Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, São Paulo, Brasil, 2011.
Disponível em http://www.teses.usp.br/teses/disponiveis/55/55134/tde-28062011-103714 (Acessado em 09/01/2014)

Coimbra, D. B.; Goularte, R. Digital video scenes identification using audiovisual features. In: *Proceedings of the XV Brazilian Symposium on Multimedia and the Web*, ACM, 2009, p. 43:1–43:4.
Disponível em http://doi.acm.org/10.1145/1858477.1858520 (Acessado em 09/01/2014)

Devasena, C. L.; Hemalatha, M. Object detection in video using lorenz information measure and discrete wavelet transform. In: *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, ICACCI '12, New York, NY, USA: ACM, 2012, p. 200–206.
Disponível em http://doi.acm.org/10.1145/2345396.2345430

Doherty, A.; Smeaton, A. Automatically segmenting lifelog data into events. In: *9th International Workshop on Image Analysis for Multimedia Interactive Services*, 2008, p. 20–23.
Disponível em http://dx.doi.org/10.1109/WIAMIS.2008.32 (Acessado em 09/01/2014)

Duthoit, C. J.; Sztynda, T.; Lal, S. K. L.; Jap, B. T.; Agbinya, J. I. Optical flow image analysis of facial expressions of human emotion: forensic applications. In: *Proceedings of the 1st International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia and Workshop*, e-Forensics '08, Brussels, Belgium: ICST, 2008, p. 5:1–5:6.
Disponível em http://dl.acm.org/citation.cfm?id=1363217.1363224 (Acessado em 09/01/2014)

Ferman, A. M.; Tekalp, A. Efficient filtering and clustering methods for temporal video segmentation and visual summarization. *Journal of Visual Communication and Image Representation*, v. 9, n. 4, p. 336–351, 1998.
Disponível em http://dx.doi.org/10.1006/jvci.1998.0402 (Acessado em 09/01/2014)

Fu, X.; Zeng, J.-x. An effective video shot boundary detection method based on the local color features of interest points. In: *Proceedings of the Second International Symposium on Electronic Commerce and Security*, Washington, DC, USA: IEEE Computer Society, 2009, p. 25–28.
Disponível em `http://dx.doi.org/10.1109/ISECS.2009.140` (Acessado em 09/01/2014)

Gargi, U.; Oswald, S.; Kosiba, D.; Devadiga, S.; Kasturi, R. Evaluation of video sequence indexing and hierarchical video indexing. In: *Proceeding on SPIE - Storage and Retrieval for Image and Video Databases*, San Jose, CA, USA, 1995.
Disponível em `http://dx.doi.org/10.1117/12.205301` (Acessado em 09/01/2014)

Geetha, P.; Narayanan, V. A survey of content-based video retrieval. *Journal of Computer Science*, v. 4, n. 6, p. 474–486, 2008.
Disponível em `http://dx.doi.org/10.3844/jcssp.2008.474.486` (Acessado em 09/01/2014)

Girisha, R.; Murali, S. Tracking humans using novel optical flow algorithm for surveillance videos. In: *Proceedings of the Fourth Annual ACM Bangalore Conference*, New York, NY, USA: ACM, 2011, p. 7:1–7:8.
Disponível em `http://dx.doi.org/10.1145/1980422.1980429` (Acessado em 09/01/2014)

Goela, N.; Wilson, K.; Niu, F.; Divakaran, A.; Otsuka, I. An svm framework for genre-independent scene change detection. In: *IEEE International Conference on Multimedia and Expo*, 2007, p. 532–535.
Disponível em `http://dx.doi.org/10.1109/ICME.2007.4284704` (Acessado em 09/01/2014)

Gonzalez, R. C.; Woods, R. E. *Digital image processing.* 3 ed. Prentice Hall, 976 p., 2007.

Graber, D. Seeing is remembering: How visuals contribute to learning from television news. *Journal of Communication*, v. 40, n. 3, p. 134–155, 1990.
Disponível em `http://dx.doi.org/10.1111/j.1460-2466.1990.tb02275.x` (Acessado em 09/01/2014)

Gu, Z.; Mei, T.; Hua, X.-S.; Wu, X.; Li, S. Ems: Energy minimization based video scene segmentation. In: *IEEE International Conference on Multimedia and Expo*, 2007, p. 520–523.

Disponível em http://dx.doi.org/10.1109/ICME.2007.4284701 (Acessado em 09/01/2014)

Günsel, B.; Ferman, A. M.; Tekalp, A. M. Temporal video segmentation using unsupervised clustering and semantic object tracking. *Journal of Electronic Imaging*, v. 7, n. 3, p. 592–604, 1998.
Disponível em http://dx.doi.org/10.1117/1.482613 (Acessado em 09/01/2014)

Hampapur, A.; Weymouth, T.; Jain, R. Digital video segmentation. In: *Proceedings of the 2nd ACM International Conference on Multimedia*, New York, NY, USA: ACM, 1994, p. 357–364.

Hanjalic, A. Shot-boundary detection: unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, v. 12, n. 2, p. 90–105, 2002.
Disponível em http://dx.doi.org/10.1109/76.988656 (Acessado em 09/01/2014)

Hong-cai, F.; Xiao-juan, Y.; Wei, M.; Cao, Y. A shot boundary detection method based on color space. In: *International Conference on E-Business and E-Government*, 2010, p. 1647–1650.
Disponível em http://dx.doi.org/10.1109/ICEE.2010.417 (Acessado em 09/01/2014)

Horn, B. K.; Schunck, B. G. *Determining optical flow.* Relatório Técnico, Cambridge, MA, USA, 1980.

Hsu, W.-M.; Chang, S.-F. Generative, discriminative, and ensemble learning on multi-modal perceptual fusion toward news video story segmentation. In: *IEEE International Conference on Multimedia and Expo*, 2004, p. 1091–1094.

Hu, W.; Xie, N.; Li, L.; Zeng, X.; Maybank, S. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics*, v. 41, n. 6, p. 797–819, 2011.
Disponível em http://dx.doi.org/10.1109/TSMCC.2011.2109710 (Acessado em 09/01/2014)

Hua, X.-S.; Zhang, D.; Li, M.; Zhang, H.-J. Performance evaluation protocol for video scene detection algorithms. In: *Workshop on Multimedia Information Retrieval, in conjunction with 10th ACM Multimedia*, 2002.

Huang, C.-R.; Chen, C.-S.; choo Chung, P. Contrast context histogram - a discriminating local descriptor for image matching. In: *Proceedings of the 18th International Conference on Pattern Recognition*, 2006, p. 53–56.

Huang, C.-R.; Chen, C.-S.; Chung, P.-C. Contrast context histogram - an efficient discriminating local descriptor for object recognition and image matching. *Pattern Recognition*, v. 41, n. 10, p. 3071–3077, 2008.
Disponível em `http://dx.doi.org/10.1016/j.patcog.2008.03.013` (Acessado em 09/01/2014)

Huang, J.; Liu, Z.; Wang, Y. Integration of audio and visual information for content-based video segmentation. In: *Proceedings of the International Conference on Image Processing*, 1998, p. 526–529.
Disponível em `http://dx.doi.org/10.1109/ICIP.1998.727252` (Acessado em 09/01/2014)

Jacobs, C. E.; Finkelstein, A.; Salesin, D. H. Fast multiresolution image querying. In: *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA: ACM, 1995, p. 277–286.
Disponível em `http://dx.doi.org/10.1145/218380.218454` (Acessado em 09/01/2014)

Jain, R.; Kasturi, R.; Schunck, B. G. *Dynamic vision.* IEEE Computer Society Press, 469-480 p., 1991.

Jeong, S. *Histogram-based color image retrieval.* Technical Report Psych221/EE362, Stanford university, 2001.
Disponível em `http://scien.stanford.edu/pages/labsite/2002/psych221/projects/02/sojeong/` (Acessado em 09/01/2014)

Jiang, H.; Helal, A.; Elmagarmid, A. K.; Joshi, A. Scene change detection techniques for video database systems. *Multimedia Systems*, v. 6, n. 3, p. 186–195, 1998.
Disponível em `http://dx.doi.org/10.1007/s005300050087` (Acessado em 09/01/2014)

Kender, J.; Yeo, B.-L. Video scene segmentation via continuous video coherence. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1998, p. 367–402.
Disponível em `http://dx.doi.org/10.1109/CVPR.1998.698632` (Acessado em 09/01/2014)

Kikukawa, T.; Kawafuchi, S. Development of an automatic summary editing system for the audio-visual resources. *Transactions on Eletronic Information*, v. E75-A, n. 2, p. 204–212, 1992.

Kim, H.; Lee, J.; Yang, J.-H.; Sull, S.; Kim, W. M.; Song, S. M.-H. Visual rhythm and shot verification. *Multimedia Tools and Applications*, v. 15, n. 3, p. 227–245, 2001. Disponível em http://dx.doi.org/10.1023/A:1012452131892 (Acessado em 09/01/2014)

Kohonen, T. The self-organizing map. *Proceedings of the IEEE*, v. 78, n. 9, p. 1464–1480, 1990.
Disponível em http://dx.doi.org/10.1109/5.58325 (Acessado em 09/01/2014)

Koprinska, I.; Carrato, S. Temporal video segmentation: A survey. *Signal Processing: Image Communication*, v. 16, n. 5, p. 477–500, 2001.
Disponível em http://dx.doi.org/10.1016/S0923-5965(00)00011-4 (Acessado em 09/01/2014)

Koumaras, H.; Gardikis, G.; Xilouris, G.; Pallis, E.; Kourtis, A. Shot boundary detection without threshold parameters. *Journal of Electronic Imaging*, v. 15, n. 2, p. 1–3, 2006. Disponível em http://dx.doi.org/10.1117/1.2199878 (Acessado em 09/01/2014)

Kristin Thompson, D. B. *Film history: An introduction.* 3 ed. New York, NY, USA: McGraw-Hill, 800 p., 2009.

Kumar, N.; Rai, P.; Pulla, C.; Jawahar, C. V. Video scene segmentation with a semantic similarity. In: *Proceedings of the 5th Indian International Conference on Artifcial Intelligence*, IICAI 2011, 2011, p. 970–981.

Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, p. 2169–2178.
Disponível em http://dx.doi.org/10.1109/CVPR.2006.68 (Acessado em 09/01/2014)

Lee, J. C.-M.; Man-Ching, D. A robust approach for camera break detection in color video sequence. In: *Proceedings of IAPR Workshop Machine Vision Applications*, Kawasaki, Japan, 1994, p. 502–505.
Disponível em http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.143.4742 (Acessado em 09/01/2014)

Lefèvre, S.; Holler, J.; Vincent, N. A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval. *Real-Time Imaging*, v. 9, n. 1, p. 73–98, 2003.
Disponível em http://dx.doi.org/10.1016/S1077-2014(02)00115-8

Li, Y.; Ming, W.; Kuo, C.-C. Semantic video content abstraction based on multiple cues. In: *IEEE International Conference on Multimedia and Expo*, ICME '01, 2001, p. 623–626.

Liu, J.; Li, M.; Liu, Q.; Lu, H.; Ma, S. Image annotation via graph learning. *Pattern Recogntion*, v. 42, n. 2, p. 218–228, 2009.
Disponível em `http://dx.doi.org/10.1016/j.patcog.2008.04.012` (Acessado em 09/01/2014)

Liu, T.; Zhang, X.; Feng, J.; Lo, K. Shot reconstruction degree: a novel criterion for key frame selection. *Pattern Recognition Letters*, v. 25, n. 12, p. 1451–1457, 2004.
Disponível em `http://dx.doi.org/10.1016/j.patrec.2004.05.020` (Acessado em 09/01/2014)

Lowe, D. Object recognition from local scale-invariant features. In: *The Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999, p. 1150–1157.

Lu, Y.; Sebe, N.; Hytnen, R.; Tian, Q. Personalization in multimedia retrieval: A survey. *Multimedia Tools and Applications*, v. 51, n. 1, p. 247–277, 2011.
Disponível em `http://dx.doi.org/10.1007/s11042-010-0621-0` (Acessado em 09/01/2014)

Lucas, B. D.; Kanade, T. An iterative image registration technique with an application to stereo vision. In: *Proceedings of the 7th International Joint Conference on Artificial intelligence*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1981, p. 674–679.
Disponível em `http://dl.acm.org/citation.cfm?id=1623264.1623280` (Acessado em 09/01/2014)

Magalhães, J. a.; Pereira, F. Using MPEG standards for multimedia customization. *Signal Processing: Image Communication*, v. 19, n. 5, p. 437–456, 2004.
Disponível em `http://dx.doi.org/10.1016/j.image.2004.02.004` (Acessado em 09/01/2014)

Manjunath, S.; Guru, D. S.; Suraj, M. G.; Harish, B. S. A non parametric shot boundary detection: an eigen gap based approach. In: *Proceedings of the 4th Annual ACM Bangalore Conference*, New York, NY, USA: ACM, 2011, p. 14:1–14:7.
Disponível em `http://doi.acm.org/10.1145/1980422.1980436` (Acessado em 09/01/2014)

Manzato, M.; Coimbra, D.; Goularte, R. An enhanced content selection mechanism for personalization of video news programmes. *Multimedia Systems*, v. 17, n. 1, p. 19–34,

10.1007/s00530-010-0204-y, 2011.
Disponível em http://dx.doi.org/10.1007/s00530-010-0204-y (Acessado em 09/01/2014)

Manzato, M. G. *Uma arquitetura de personalização de conteúdo baseada em anotações do usuário.* Tese (doutorado em ciência da computação e matemática computacional), Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, São Paulo, Brasil, 2011.
Disponível em http://www.teses.usp.br/teses/disponiveis/55/55134/tde-11042011-160836 (Acessado em 09/01/2014)

Manzato, M. G.; Coimbra, D. B.; Goularte, R. Multimedia content personalization based on peer-level annotation. In: *Proceedings of the 7th European Conference on European Interactive Television Conference*, EuroITV '09, New York, NY, USA: ACM, 2009a, p. 57–66.
Disponível em http://doi.acm.org/10.1145/1542084.1542096

Manzato, M. G.; Goularte, R. Shot boundary detection based on intelligent systems. In: *Proceedings of the XIII Brazilian Symposium on Multimedia and the Web*, WebMedia '07, SBC, 2007, p. 190–197.

Manzato, M. G.; Goularte, R. Peersommender: A peer-level annotation based approach for multimedia recommendation. *Journal of Information and Data Management*, v. 1, n. 2, p. 277–292, 2010.
Disponível em http://dblp.uni-trier.de/db/journals/jidm/jidm1.html#ManzatoG10 (Acessado em 09/01/2014)

Manzato, M. G.; Goularte, R. A multimedia recommender system based on enriched user profiles. In: *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, New York, NY, USA: ACM, 2012, p. 975–980.
Disponível em http://doi.acm.org/10.1145/2245276.2245466

Manzato, M. G.; Macedo, A. A.; Goularte, R. Evaluation of video news classification techniques for automatic content personalization. *International Journal of Advanced Media and Communication*, v. 3, n. 4, p. 383–403, 2009b.
Disponível em http://dl.acm.org/citation.cfm?id=1628257.1628260 (Acessado em 09/01/2014)

Marques, O. *Practical image and video processing using matlab.* Wiley, IEEE Press, 696 p., 2011.

Marques Filho, O.; Vieira Neto, H. *Processamento digital de imagens.* 1 ed. Rio de Janeiro, RJ, Brasil: Brasport, 331 p., 1999.

Mood, A. M.; Graybill, F. A.; Boes, D. C. *Introduction to the theory of statistics.* 3 ed. New York, NY, USA: McGraw Hill, 480 p., 1974.

Mukherjee, D.; Das, S.; Saha, S. Key frame estimation in video using randomness measure of feature point pattern. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 17, n. 5, p. 612–620, 2007.
Disponível em `http://dx.doi.org/10.1109/TCSVT.2007.895353` (Acessado em 09/01/2014)

Neumann, B. Optical flow. *SIGGRAPH Computer Graphics*, v. 18, n. 1, p. 17–19, 1984.
Disponível em `http://dx.doi.org/10.1145/988525.988528` (Acessado em 09/01/2014)

Ogawa, A.; Takahashi, T.; Ide, I.; Murase, H. Cross-lingual retrieval of identical news events by near-duplicate video segment detection. In: *Proceedings of the 14th International Conference on Advances in Multimedia Modeling*, Berlin, Heidelberg: Springer-Verlag, 2008, p. 287–296.
Disponível em `http://dl.acm.org/citation.cfm?id=1785794.1785826` (Acessado em 09/01/2014)

Papadopoulos, G. T.; Mezaris, V.; Kompatsiaris, I.; Strintzis, M. G. Accumulated motion energy fields estimation and representation for semantic event detection. In: *Proceedings of the International Conference on Content-based Image and Video Retrieval*, New York, NY, USA: ACM, 2008, p. 221–230.
Disponível em `http://doi.acm.org/10.1145/1386352.1386385` (Acessado em 09/01/2014)

Pass, G.; Zabih, R. Comparing images using joint histograms. *Multimedia Systems*, v. 7, n. 3, p. 234–240, 1999.
Disponível em `http://dx.doi.org/10.1007/s005300050125` (Acessado em 09/01/2014)

Pfeiffer, S.; Lienhart, R.; Efflsberg, W. Scene determination based on video and audio features. *Multimedia Tools and Applications*, v. 15, n. 1, p. 59–81, 2001.
Disponível em `http://dx.doi.org/10.1023/A:1011315803415` (Acessado em 09/01/2014)

Pickering, M.; Rüger, S. Evaluation of key frame-based retrieval techniques for video. *Computer Vision and Image Understanding*, v. 92, n. 2-3, p. 217–235, 2003.
Disponível em `http://oro.open.ac.uk/8045/` (Acessado em 09/01/2014)

Pimentel Filho, C. A. F.; Santos, C. A. S.; Buck, T. A. Integração de métodos baseados em diferença de quadros para sumarização do conteúdo de vídeos. In: *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, New York, NY, USA: ACM, 2008, p. 85–88.
Disponível em `http://dx.doi.org/10.1145/1809980.1810003` (Acessado em 09/01/2014)

Poulisse, G.; Moens, M. Unsupervised scene detection in olympic video using multi-modal chains. In: *9th International Workshop on Content-Based Multimedia Indexing*, 2011, p. 103–108.
Disponível em `http://dx.doi.org/10.1109/CBMI.2011.5972529` (Acessado em 09/01/2014)

Punitha, P.; Jose, J. M. Shot boundary detection based on eigen coefficients and small eigen value. In: *Proceedings of the 4th International Conference on Semantic and Digital Media Technologies: Semantic Multimedia*, Berlin, Heidelberg: Springer-Verlag, 2009, p. 126–136.
Disponível em `http://dx.doi.org/10.1007/978-3-642-10543-2_13` (Acessado em 09/01/2014)

Rasheed, Z.; Shah, M. Scene detection in hollywood movies and tv shows. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003, p. 343–348.
Disponível em `http://dx.doi.org/10.1109/CVPR.2003.1211489` (Acessado em 09/01/2014)

Rasheed, Z.; Shah, M. Detection and representation of scenes in videos. *IEEE Transactions on Multimedia*, v. 7, n. 6, p. 1097–1105, 2005.
Disponível em `http://dx.doi.org/10.1109/TMM.2005.858392` (Acessado em 09/01/2014)

Rathod, G. I.; Nikam, D. A. An algorithm for shot boundary detection and key frame extraction using histogram difference. *International Journal of Emerging Technology and Advanced Engineering*, v. 3, n. 8, p. 155–163, 2013.
Disponível em `http://ijetae.com/Volume3Issue8.html` (Acessado em 03/10/2013)

Richardson, I. E. *The h.264 advanced video compression standard.* 2 ed. Chichester: John Wiley & Sons Ltd, 2010.

Rijsbergen, C. G. *Information retrieval.* 2 ed. London: Butterworths, 224 p., 1979.

Sakarya, U.; Telatar, Z. Graph partition based scene boundary detection. In: *5th International Symposium on Image and Signal Processing and Analysis*, 2007, p. 544–549. Disponível em `http://dx.doi.org/10.1109/ISPA.2007.4383752` (Acessado em 09/01/2014)

Sakarya, U.; Telatar, Z. Video scene detection using dominant sets. In: *15th IEEE International Conference on Image Processing*, 2008, p. 73–76. Disponível em `http://dx.doi.org/10.1109/ICIP.2008.4711694` (Acessado em 09/01/2014)

Sakarya, U.; Telatar, Z. Video scene detection using graph-based representations. *Signal Processing: Image Communication*, v. 25, n. 10, p. 774–783, 2010. Disponível em `http://dx.doi.org/10.1016/j.image.2010.10.001` (Acessado em 09/01/2014)

Salomon, D.; Motta, G.; Bryant, D. *Data compression: The complete reference.* 4 ed. New York, NY, USA: Springer, 2006.

Sarris, V. [Max Wertheimer in Frankfurt-on the beginnings and developmental crisis of Gestalt psychology. Initial studies of motion perception (1910-1912)]. *Zeitschrift für Psychologie mit Zeitschrift für angewandte Psychologie*, v. 195, n. 3, p. 283–310, 1987. Disponível em `http://www.ncbi.nlm.nih.gov/pubmed/2895552` (Acessado em 09/01/2014)

Sawai, K.; Takahashi, T.; Deguchi, D.; Ide, I.; Murase, H. Scene segmentation of wedding party videos by scenario-based matching with example videos. In: *Proceedings of the 19th ACM International Conference on Multimedia*, New York, NY, USA: ACM, 2011, p. 1545–1548. Disponível em `http://doi.acm.org/10.1145/2072298.2072061` (Acessado em 09/01/2014)

Shahraray, B. Scene change detection and content-based sampling of video sequences. In: *Proceedings of the Storage and Retrieval for Image and Video Databases*, 1995, p. 2–13. Disponível em `http://dx.doi.org/10.1117/12.206348` (Acessado em 09/01/2014)

Sivic, J.; Everingham, M.; Zisserman, A. Person spotting: Video shot retrieval for face sets. In: *Proceedings of the 4th International Conference on Image and Video Retrieval*, CIVR '05, Springer, 2005, p. 226–236.
Disponível em `http://dx.doi.org/10.1007/11526346_26` (Acessado em 09/01/2014)

Smeaton, A. F.; Over, P.; Doherty, A. R. Video shot boundary detection: Seven years of trecvid activity. *Computer Vision and Image Understanding*, v. 114, n. 4, p. 411 – 418, special issue on Image and Video Retrieval Evaluation, 2010.
Disponível em `http://dx.doi.org/10.1016/j.cviu.2009.03.011` (Acessado em 09/01/2014)

Smeaton, A. F.; Over, P.; Kraaij, W. Evaluation campaigns and trecvid. In: *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, New York, NY, USA: ACM, 2006, p. 321–330.
Disponível em `http://dx.doi.org/10.1145/1178677.1178722` (Acessado em 09/01/2014)

Smith, A. R. Color gamut transform pairs. *SIGGRAPH Computer Graphics*, v. 12, n. 3, p. 12–19, 1978.
Disponível em `http://doi.acm.org/10.1145/965139.807361` (Acessado em 09/01/2014)

Snoek, C. G. M.; Worring, M.; van Gemert, J.; Geusebroek, J.-M.; Koelma, D.; Nguyen, G. P.; de Rooij, O.; Seinstra, F. Mediamill: exploring news video archives based on learned semantics. In: *Proceedings of the 13th annual ACM International Conference on Multimedia*, New York, NY, USA: ACM, 2005, p. 225–226.
Disponível em `http://doi.acm.org/10.1145/1101149.1101188` (Acessado em 09/01/2014)

Song, Y.; Ogawa, T.; Haseyama, M. Mcmc-based scene segmentation method using structure of video. In: *International Symposium on Communications and Information Technologies*, ISCIT '10, 2010, p. 862 –866.
Disponível em `http://dx.doi.org/10.1109/ISCIT.2010.5665107` (Acessado em 09/01/2014)

Souza, T. T.; Goularte, R. Video shot representation based on histograms. In: *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, New York, NY, USA: ACM, 2013, p. 961–966.
Disponível em `http://doi.acm.org/10.1145/2480362.2480547` (Acessado em 09/01/2014)

Sowjanya, P. S.; Mishra, R. Video shot boundary detection - comparision of color histogram and gist method. *International Journal of Research Engineering and Applied Sciences*, v. 1, n. 1, p. 1336–1350, 2012.

Sujatha, C.; Mudenagudi, U. A study on keyframe extraction methods for video summary. In: *International Conference on Computational Intelligence and Communication Networks*, CICN '11, 2011, p. 73–77.

Sundaram, H.; Chang, S. F. Video scene segmentation using video and audio features. In: *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2000, p. 1145–1148.
Disponível em `http://dx.doi.org/10.1109/ICME.2000.871563` (Acessado em 09/01/2014)

Surlea, C.; Kurugollu, F.; Milligan, P.; Ong, S. Foetal motion classification using optical flow displacement histograms. In: *Proceedings of the 4th International Symposium on Applied Sciences in Biomedical and Communication Technologies*, New York, NY, USA: ACM, 2011, p. 156:1–156:5.
Disponível em `http://doi.acm.org/10.1145/2093698.2093854` (Acessado em 09/01/2014)

Swain, M. Interactive indexing into image databases. In: Niblack, C. W., ed. *Storage and Retrieval for Image and Video Databases*, 1993, p. 95–103.
Disponível em `http://dx.doi.org/10.1117/12.143659` (Acessado em 09/01/2014)

Swanberg, D.; Shu, C.-F.; Jain, R. Knowledge-guided parsing in video databases. In: Niblack, C. W., ed. *Proceedings of the Storage and Retrieval for Image and Video Databases*, 1993, p. 13–24.
Disponível em `http://dx.doi.org/10.1117/12.143647` (Acessado em 09/01/2014)

Takimoto, M.; Satoh, S.; Sakauchi, M. Identification and detection of the same scene based on flash light patterns. In: *IEEE International Conference on Multimedia and Expo*, 2006, p. 9–12.
Disponível em `http://dx.doi.org/10.1109/ICME.2006.262501` (Acessado em 09/01/2014)

Tapu, R.; Zaharia, T. High level video temporal segmentation. In: *Proceedings of the 7th International Conference on Advances in Visual Computing*, Springer Berlin / Heidelberg, 2011, p. 224–235.
Disponível em `http://dx.doi.org/10.1007/978-3-642-24028-7_21` (Acessado em 09/01/2014)

Tavanapong, W.; Zhou, J. Shot clustering techniques for story browsing. *IEEE Transactions on Multimedia*, v. 6, n. 4, p. 517–527, 2004.
Disponível em `http://dx.doi.org/10.1109/TMM.2004.830810` (Acessado em 09/01/2014)

Toffler, A. *Future shock*. 1 ed. Bantam, 576 p., 1984.

Trojahn, T. H.; Goularte, R. *Segmentação de vídeo baseada em coerência entre tomadas*. Relatório Técnico, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos: Biblioteca Achilli Bassi, 2013.

Ullman, S.; Vidal-Naquet, M.; Sali, E. Visual features of intermediate complexity and their use in classification. *Nature neuroscience*, v. 5, n. 7, p. 682–687, 2002.
Disponível em `http://dx.doi.org/10.1038/nn870` (Acessado em 09/01/2014)

Uras, S.; Girosi, F.; Verri, A.; Torre, V. A computational approach to motion perception. *Biological Cybernetics*, v. 60, n. 2, p. 79–87, 1988.
Disponível em `http://dx.doi.org/10.1007/BF00202895` (Acessado em 09/01/2014)

Viola, P.; Jones, M. Robust real-time face detection. In: *Proceedings of the 8th IEEE International Conference on Computer Vision*, 2001, p. 747.
Disponível em `http://dx.doi.org/10.1109/ICCV.2001.937709` (Acessado em 09/01/2014)

Wang, C.; Jing, F.; Zhang, L.; Zhang, H.-J. Scalable search-based image annotation. *Multimedia Systems*, v. 14, n. 4, p. 205–220, 2008a.
Disponível em `http://dx.doi.org/10.1007/s00530-008-0128-y` (Acessado em 09/01/2014)

Wang, J.; Tian, X.; Yang, Y.; Zha, Z.-J.; Hua, X.-S. Optimized video scene segmentation. In: *IEEE International Conference on Multimedia and Expo*, 2008b, p. 301–304.
Disponível em `http://dx.doi.org/10.1109/ICME.2008.4607431` (Acessado em 09/01/2014)

Wang, Y. Foreground and shadow detection based on conditional random field. In: *Proceedings of the 12th international conference on Computer analysis of images and patterns*, CAIP'07, Berlin, Heidelberg: Springer-Verlag, 2007, p. 85–92.
Disponível em `http://dl.acm.org/citation.cfm?id=1770904.1770917` (Acessado em 09/01/2014)

Wikipedia Zoetrope. Website, 2013.
Disponível em `http://en.wikipedia.org/wiki/Zoetrope` (Acessado em 09/01/2014)

Wu, B.; Jiang, X.; Sun, T.; Zhang, S.; Chu, X.; Shen, C.; Fan, J. A novel horror scene detection scheme on revised multiple instance learning model. In: Lee, K.-T.; Tsai, W.-H.; Liao, H.-Y.; Chen, T.; Hsieh, J.-W.; Tseng, C.-C., eds. *Advances in Multimedia Modeling*, v. 6524 de *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, p. 359–370, 2011.
Disponível em `http://dx.doi.org/10.1007/978-3-642-17829-0_34` (Acessado em 09/01/2014)

Xiang-wei, L.; Ming-xin, Z.; Lang-tao, Z.; Jin-hong, X.; Ya-lin, Z. Effective shot segmentation method based on rough sets in video compressed domain. In: *Proceedings of the World Congress on Computer Science and Information Engineering*, Los Angeles, CA, USA, 2009, p. 421–425.
Disponível em `http://dx.doi.org/10.1109/CSIE.2009.304`

Xie, L.; Xu, P.; Chang, S.-F.; Divakaran, A.; Sun, H. Structure analysis of soccer video with domain knowledge and hidden markov models. *Pattern Recognition Letters*, v. 25, n. 7, p. 767–775, 2004.
Disponível em `http://dx.doi.org/10.1016/j.patrec.2004.01.005` (Acessado em 09/01/2014)

Yeung, M.; Yeo, B.-L.; Liu, B. Segmentation of video by clustering and graph analysis. *Computer Vision and Image Understanding*, v. 71, n. 1, p. 94–109, 1998.
Disponível em `http://dx.doi.org/10.1006/cviu.1997.0628` (Acessado em 09/01/2014)

Yu, H.; Su, B.; Lu, H.; Xue, X. News video retrieval by learning multimodal semantic information. In: *Advances in Visual Information Systems*, v. 4781 de *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, p. 403–414, 2007a.
Disponível em `http://dx.doi.org/10.1007/978-3-540-76414-4_39`

Yu, J.; Tian, B.; Tang, Y. Video segmentation based on shot boundary coefficient. In: *Proceedings of the 2nd International Conference on Pervasive Computing and Applications*, 2007b, p. 630–635.
Disponível em `http://dx.doi.org/10.1109/ICPCA.2007.4365519` (Acessado em 09/01/2014)

Yuan, J.; Wang, H.; Xiao, L.; Zheng, W.; Li, J.; Lin, F.; Zhang, B. A formal study of shot boundary detection. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 17, n. 2, p. 168–186, 2007.
Disponível em `http://dx.doi.org/10.1109/TCSVT.2006.888023` (Acessado em 09/01/2014)

Yuan, J.; Wei, B.; Lu, W.; Wang, L. A new video text detection method. In: *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, New York, NY, USA: ACM, 2011, p. 359–362.
Disponível em `http://doi.acm.org/10.1145/1998076.1998142`

Zabih, R.; Miller, J.; Mai, K. A feature-based algorithm for detecting and classifying production effects. *Multimedia Systems*, v. 7, n. 2, p. 119–128, 1999.
Disponível em `http://dx.doi.org/10.1007/s005300050115` (Acessado em 09/01/2014)

Zhai, Y.; Shah, M. Determining structure in continuously recorded videos. In: *Proceedings of the 13th annual ACM International Conference on Multimedia*, New York, NY, USA: ACM, 2005a, p. 495–498.
Disponível em `http://dx.doi.org/10.1145/1101149.1101260` (Acessado em 09/01/2014)

Zhai, Y.; Shah, M. A general framework for temporal video scene segmentation. In: *Proceedings of the 10th IEEE International Conference on Computer Vision*, 2005b, p. 1111–1116.
Disponível em `http://dx.doi.org/10.1109/ICCV.2005.6` (Acessado em 09/01/2014)

Zhai, Y.; Shah, M. Video scene segmentation using markov chain monte carlo. *IEEE Transactions on Multimedia*, v. 8, n. 4, p. 686–697, 2006.
Disponível em `http://dx.doi.org/10.1109/TMM.2006.876299` (Acessado em 09/01/2014)

Zhang, C.; Wang, W. A robust and efficient shot boundary detection approach based on fisher criterion. In: *Proceedings of the 20th ACM international Conference on Multimedia*, New York, NY, USA: ACM, 2012, p. 701–704.
Disponível em `http://doi.acm.org/10.1145/2393347.2396291`

Zhang, H.; Kankanhalli, A.; Smoliar, S. W. Automatic partitioning of full-motion video. *Multimedia Systems*, v. 1, n. 1, p. 10–28, 1993.
Disponível em `http://dx.doi.org/10.1007/BF01210504` (Acessado em 09/01/2014)

Zhang, L.; Zhu, Z.; Zhao, Y. Robust commercial detection system. In: *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2007, p. 587–590.
Disponível em `http://dx.doi.org/10.1109/ICME.2007.4284718` (Acessado em 09/01/2014)