SUMARIZAÇÃO AUTOMÁTICA DE *RUSHES VIDEOS* BASEADA EM CARACTERÍSTICAS ESPACIAIS E ESPAÇO-TEMPORAIS

TIAGO OLIVEIRA. CUNHA

SUMARIZAÇÃO AUTOMÁTICA DE *RUSHES VIDEOS* BASEADA EM CARACTERÍSTICAS ESPACIAIS E ESPAÇO-TEMPORAIS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Arnaldo de Albuquerque Araújo
Co-orientadora: Gisele Lobo Pappa

Belo Horizonte
Abril de 2011

Cunha, Tiago Oliveira.
C972s Sumarização automática de *rushes videos* baseada em características espaciais e espaço-temporais / Tiago Oliveira. Cunha. — Belo Horizonte, 2011
xxii, 66 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Arnaldo de Albuquerque Araújo

Co-orientador: Gisele Lobo Pappa

1. Sumarização de vídeos. 2. Características locais. 3. Agrupamento. 4. rushes videos. I. Título.

CDU 519.6*84 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Sumarização Automática de vídeos baseada em características espaciais e espaço-temporais

TIAGO OLIVEIRA CUNHA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. ARNALDO DE ALBUQUERQUE ARAÚJO - Orientador
Departamento de Ciência da Computação - UFMG

PROFA. GISELE LOBO PAPPA - Co-orientadora
Departamento de Ciência da Computação - UFMG

PROF. NEUCIMAR JERÔNIMO LEITE
Instituto de Computação - UNICAMP

PROF. SILVIO JAMIL FERZOLI GUIMARÃES
Instituto de Informática - PUC/MG

Belo Horizonte, 03 de junho de 2011.

Agradecimentos

Agradeço primeiro a Deus por mais uma etapa concluída na minha vida.

A meus pais por todo o esforço que fizeram e ainda fazem por mim, estando ao meu lado em todos os momentos. A minha irmã por toda ajuda que sempre tem a me oferece.

Aos colegas do NPDI, especialmente Júlia, Marcelo e Sandra, que estiveram presentes durante todo o período do mestrado e foram muito importantes na parte pessoal e profissional.

A Fillipe que foi um grande amigo e colega de trabalho durante esses 2 anos de mestrado.

Ao Flávio Souza e Elerson pela ajuda muito importante com partes do código.

A meu orientador, Arnaldo de Albuquerque Araújo pela orientação e acompanhamento durante o trabalho.

Um agradecimento mais que especial para a minha co-orientadora Gisele Lobo Pappa, que contribuiu de forma decisiva para a conclusão desse trabalho, sempre estando a meu lado e se tornando mais que uma orientadora, mas sim uma amiga pessoal.

Aos inúmeros amigos do DCC, em especial os companheiros do RFC, Rafael Colares, Rafael Santin e Zilton.

Aos meus colegas de república, Flávio, André e Giovanni pela ótima convivência durante mais de 2 anos morando juntos.

Aos amigos que fiz em Belo horizonte, em especial Renata Onety e Luisa Vieira.

Por fim a todos que colaboraram direta ou indiretamente para a conclusão desse trabalho, meus sinceros agradecimentos.

"Se você já construiu castelos no ar, não tenha vergonha deles.
Estão onde devem estar. Agora, dê-lhes alicerces."
(Henry David Thoreau)

Resumo

Sumarização automática de vídeos é uma tarefa que permite a síntese do conteúdo de um vídeo preservando as sequências mais importantes ou mais representativas.

Este trabalho trata da geração de resumos para um tipo específico de vídeo, os *rushes videos*. *Rushes videos* são vídeos não editados produzidos durante a gravação de uma produção videográfica. Eles apresentam uma estrutura muito específica, com gravações de uma cena em vários ângulos, comentários do diretor, claquetes, barras de cor, etc. Esses vídeos são sumarizados de forma diferente dos vídeos tradicionais devido a sua estrutura. Sumarizar *rushes videos* pode trazer benefícios para o processo de edição, que é realizado manualmente pelo editor de filme. Além disso, outro foco do trabalho é a utilização e comparação de diversas caracterizações do vídeo.

Neste trabalho é apresentada uma nova abordagem para a resolução do problema de geração automática de resumos dinâmicos de *rushes videos*. O método se beneficia do uso de características espaciais e espaço-temporais associadas a uma técnica de histograma de palavras visuais e agrupamento para a formação dos resumos. A avaliação dos resumos foi feita por usuários e os resultados comparados com trabalhos encontrados na literatura. Os resultados mostraram que a metodologia proposta alcançou qualidade estatisticamente similar em relação às abordagens comparadas.

Palavras-chave: Sumarização de vídeos, Características locais, Agrupamento, aprendizado multivisão.

Abstract

Automatic video summarization is a task that allows the synthesis of the video content while preserving the most important or most representative sequences.

This work deals with the summarization process of a specific video type: rushes videos. Rushes videos are the raw material recorded during a video production. They present a very specific structure, with several takes of scenes in different angles, director's commentaries, clapperboads, color bars, etc. These videos are summarized in a different way than traditional videos due their specific structure. Summarizing rushes videos can bring benefits to the editing process, which is performed manually by the film editor. In addition, another focus of this work is the use and comparison of different video representations.

We present a new approach for solving the problem of automatic generation of dynamic video summaries of rushes videos. The method benefits from the use of clustering, spatial and spatiotemporal features associated with a bag-of-visual-features technique for summaries generation. The evaluation was done by users and the results compared with the literature. The proposed methodology achieved statistically similar quality in relation to the approaches compared.

Keywords: Video summarization, Local Features, Clustering, Multiview Learning, Segmentation.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Os avanços em técnicas de compressão, a diminuição no custo de equipamentos para aquisição e armazenamento de vídeo e a disponibilidade de meios de transmissão de dados em alta velocidade têm facilitado a forma como os vídeos são criados, armazenados e distribuídos. Como consequência, os vídeos passaram a ser utilizados em várias aplicações. Porém, as técnicas de gerenciamento de dados textuais disponíveis mostraram-se ineficientes para dados tão distintos dos "tradicionais".

Por estes motivos, a pesquisa e o desenvolvimento de novas tecnologias tornaram-se necessários para um gerenciamento mais eficiente de vídeos. Dentre as diversas áreas possíveis de pesquisa, a sumarização automática de vídeos é uma etapa essencial para inúmeras aplicações, tais como indexação, navegação e recuperação por conteúdo ([Truong & Venkatesh, 2007]).

Sumarização automática de vídeos é uma tarefa que permite a síntese do conteúdo de um vídeo, preservando as sequências mais importantes ou mais representativas. Nesse processo, o conteúdo do vídeo é analisado, a sua estrutura é identificada e então os segmentos mais relevantes são selecionados (Dumont & Mérialdo [2010]).

Os resumos gerados a partir dos métodos de sumarização automática de vídeos, podem ser classificados em duas categorias principais: estáticos ou dinâmicos. A primeira categoria consiste na extração de um conjunto de quadros-chave do vídeo original. Já a segunda categoria coleta um conjunto de segmentos do vídeo original. Esses segmentos são unidos por transições abruptas ou graduais, formando um vídeo clip com uma duração significativamente menor que a do vídeo original. Uma vantagem dos resumos dinâmicos em relação aos resumos estáticos é a possibilidade de incluir elementos de áudio e movimentos, realçando tanto a expressividade quanto a informação do resumo.

A pesquisa em torno da geração de resumos dinâmicos é relativamente nova.

É uma tarefa difícil, pois a natureza multimodal dos vídeos (áudio, texto, imagens estáticas e imagens em movimento) incorpora uma grande quantidade de conceitos semânticos. Assim um dos grandes desafios dessa área é o desenvolvimento de técnicas para abstrair semântica útil e intuitiva dos vídeos, que esteja de acordo com a compreensão e entendimento do usuário (Money & Agius [2008]). Outro ponto importante a ser salientado é a falta de um modelo padrão para a avaliação dos resumos, assim como a falta de base de dados que possuam anotações, o que torna a comparação dos trabalhos produzidos na literatura difícil, se não impossível.

A abordagem proposta no presente trabalho se concentra na geração de resumos dinâmicos de *rushes videos*. *Rushes videos* consistem de material não editado que contém várias gravações de cenas de um vídeo. Por exemplo, durante a gravação de uma produção videográfica são gravadas cenas de acordo com o *script*. Normalmente, uma cena é gravada várias vezes, com variações indicadas pelo diretor da produção. Além disso, *rushes videos* possuem uma série de dados que são utilizados apenas para fins de marcação e separação entre gravações, como por exemplo: segmentos de padrões de teste para calibrar as cores da câmera, trechos com barras de cor e sequências que possuem claquetes que identificam a gravação e o número da cena na gravação.

Rushes são potencialmente muito valiosos, mas largamente inexplorados. Somente a equipe de produção original sabe o que os rushes contêm, e os metadados são geralmente muito limitados, incluindo apenas poucas informações de indexação, tais como pelo programa, departamento, nome e data. De 20 a 40 horas de *rushes* podem ser gravadas para cada hora de programação produzida (Over et al. [2007]). Espera-se que a capacidade de resumir esses *rushes* possa contribuir significativamente para o processo de edição das produções, uma vez que esse processo é realizado manualmente pelo editor de filme.

Embora muitas técnicas tenham sido propostas para resumir automaticamente o conteúdo de vídeos em geral, a estrutura específica dos *rushes videos* requer uma adaptação destas técnicas e, às vezes, o desenvolvimento de novas abordagens para uma sumarização eficaz (Dumont & Mérialdo [2010]).

Na literatura, são encontradas diferentes técnicas para geração de resumos a partir de *rushes videos* (Dumont & Mérialdo [2010], Pan et al. [2007], Le & Satoh [2007], Beran et al. [2007], Chen et al. [2008], Detyniecki & Marsala [2007]), onde a maioria se baseia em técnicas de agrupamento. Esse tipo de abordagem se fundamenta na geração dos resumos a partir da remoção de redundância entre os segmentos do vídeo. Então, os segmentos semelhantes do vídeo devem ser caracterizados de maneira que permitam uma efetiva comparação entre si. São utilizadas diversas maneiras de caracterização, pois não há um consenso sobre quais as melhores características a serem

usadas para essa tarefa.

Este trabalho apresenta uma abordagem para a resolução do problema de geração automática de resumos dinâmicos de *rushes videos*. O método se beneficia do uso de características espaciais e espaço-temporais associadas a uma técnica de histograma de palavras visuais e agrupamento para a formação dos resumos. A representação por histogramas de palavras visuais espaço-temporais tem sido utilizada na literatura em vários cenários de detecção e classificação de padrões, alcançando bons resultados, devido a sua robustez a uma série de transformações na imagem e oclusão. Portanto, essa técnica mostra-se adequada para a análise em alto nível de imagens/vídeos. Entretanto, pelo levantamento bibliográfico realizado, não foi encontrado nenhum trabalho sobre sumarização de vídeos que utilize essa representação.

1.1 Objetivos

O objetivo geral deste trabalho é desenvolver uma abordagem eficaz para sumarização automática de vídeos. A abordagem proposta é baseada na utilização de características espaciais e espaço-temporais associadas a uma técnica de histogramas de palavras visuais e agrupamento. Estes resumos devem ser suficientemente representativos e concisos.

Os objetivos específicos consistem em:

1. Utilizar uma técnica de histograma de palavras visuais em um problema classificação não-supervisionada;

2. Selecionar uma métrica eficaz e eficiente para a seleção dos segmentos que irão compor o resumo;

3. Comparar a qualidade dos resumos gerados a partir de características espaciais e espaço-temporais.

4. Analisar a qualidade dos resumos formados a partir da união dos resumos gerados por características espaciais e espaço-temporais.

5. Selecionar um modelo de avaliação que permita quantificar a qualidade do resumo e comparar os resultados com os encontrados na literatura.

1.2 Justificativa e Relevância

Atualmente, vídeos são utilizados em diversas aplicações, principalmente aplicações *Web*, pois são uma maneira fácil e agradável de obter informação e prover entretenimento. Nos últimos anos surgiram aplicações específicas para compartilhamento de vídeos (*Youtube*[1],*Dailymotion*[2], etc.), o que facilitou o acesso do usuário a grandes quantidades de informação. Porém, as soluções para gerenciamento não acompanharam o crescimento da oferta de informação de vídeo, tornando a análise das informações de grandes quantidades de vídeos uma tarefa complexa e demorada.

Assim sendo, resumos de vídeos aparecem como uma solução que permite uma identificação prévia sobre o conteúdo dos vídeos, diminuindo a quantidade de tempo gasta em relação à análise do conteúdo do vídeo original. O sumário possibilita ao usuário julgar se o vídeo é realmente relevante para o fim desejado, auxiliando na tomada de decisões.

As principais contribuições do trabalho proposto se concentram na definição de um modelo capaz de gerar automaticamente resumos dinâmicos de *rushes videos*. A metodologia proposta incorpora vantagens observadas na literatura relacionada, agregando em único método técnicas de comprovada eficácia na geração de resumos dinâmicos.

Uma outra contribuição é a análise do comportamento de técnicas que são o estado da arte no reconhecimento de padrões em detecção e reconhecimento de objetos e ações humanas, quando aplicadas ao problema de sumarização dinâmica de vídeos.

1.3 Organização da dissertação

Este trabalho está organizado da seguinte maneira: No Capítulo 2, é apresentada uma revisão teórica sobre o processo de sumarização automática de vídeos, especificamente voltada para a geração de resumos dinâmicos. O Capítulo 3 apresenta a metodologia desenvolvida neste trabalho e metodologia de avaliação utilizados para estimar a qualidade dos resumos produzidos. No Capítulo 4, é apresentada a base de dados utilizada nos experimentos, os resultados obtidos e a análise destes. As conclusões gerais e os trabalhos futuros são expostos no Capítulo 5 deste trabalho.

[1]www.youtube.com
[2]www.dailymotion.com

Capítulo 2

Sumarização dinâmica de vídeos

A geração de resumos dinâmicos é uma área relativamente nova de pesquisa, e geralmente requer grande análise do conteúdo do vídeo (Truong & Venkatesh [2007]).

A técnica mais simples para a geração de resumos dinâmicos é a subamostragem, que seleciona segmentos de tamanho pré-definido em intervalos fixos. Um segundo tipo de técnicas simplesmente aumenta a taxa de quadros por segundo do vídeo inteiro, seja uniformemente ou baseado em análise dinâmica do conteúdo do vídeo. Na primeira técnica pode haver problemas de redundância no resumo, umas vez que não há nenhuma etapa de eliminação de conteúdo similar. Um outro problema é o fato de não haver nenhum critério de importância na seleção dos segmentos que irão compor o resumo. A segunda técnica pode causar perda de coerência e desconforto ao espectador, pois com a aceleração do vídeo algumas cenas se tornam difíceis de compreender.

Outra abordagem é gerar o resumo dinâmico a partir do resumo estático, pois uma vez que se tem conhecimento sobre os quadros-chave do vídeo, pode-se definir uma vizinhança de quadros para a formação de segmentos contínuos, que são unidos para formar o resumo dinâmico, assim como feito por Wu et al. [2000].

Esse capítulo irá tratar das técnicas de geração de resumos dinâmicos que mantêm a mesma taxa de quadros por segundo do vídeo original, e que definem um critério de relevância/importância para a escolha dos segmentos que irão compor o resumo. A Figura 2.1 mostra a caracterização das técnicas de geração de resumos dinâmicos empregadas na literatura. Essa caracterização foi proposta por Truong & Venkatesh [2007], e nela são identificados cinco aspectos principais na geração de resumos dinâmicos: o gênero dos vídeos a serem sumarizados, a natureza da determinação da duração do resumo (descritos na Seção 2.1), as etapas do processo (Seção 2.2), o mecanismo de sumarização(Seção 2.4) e as características utilizadas (Seção 2.5). As próximas seções detalham cada um desses passos, além de abordar as técnicas de avaliação de resumos

dinâmicos.

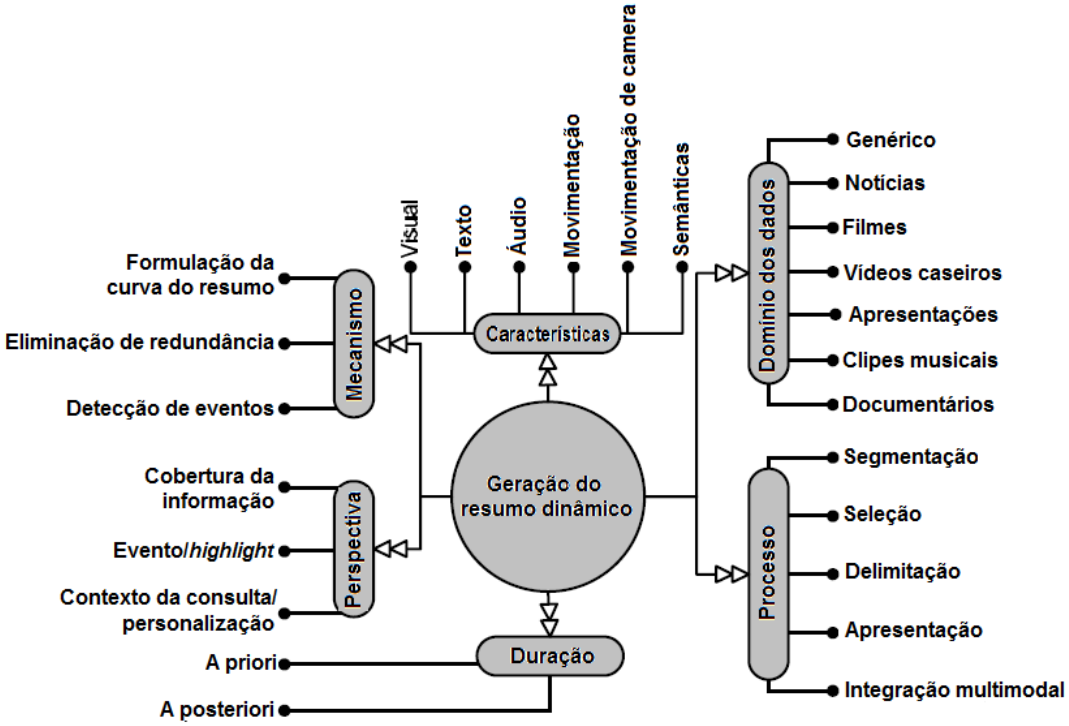

Figura 2.1. Atributo das técnicas de sumarização dinâmica (Truong & Venkatesh [2007]).

2.1 Definições para a construção do sumário dinâmico

Existem duas decisões que devem ser tomadas antes do início da construção do sumário: a determinação da duração do resumo e o domínio dos dados a serem trabalhados. Essas decisões irão influenciar diretamente no tipo de perspectiva e no tipo de mecanismo utilizado na formação do resumo.

Em relação ao **domínio dos dados a serem trabalhados** a técnica utilizada para a geração do resumo preferencialmente deve tirar proveito das características do tipo de vídeo a ser sumarizado.

Uma técnica é considerada genérica se ela pode ser utilizada para a geração de resumos de qualquer tipo de vídeo. Porém, a maioria dos trabalhos encontrados na literatura são dependentes do domínio em que se encontram os vídeos, pois a exploração de características específicas de cada tipo de vídeo pode tornar essas técnicas mais

eficazes, alcançando melhores resultados. As técnicas classificadas como genéricas geralmente são baseadas em eliminação de redundância.

Na literatura são encontrados trabalhos para os seguintes tipos de vídeos: esportes (Tjondronegoro et al. [2003], Tjondronegoro et al. [2004]), notícias (Lie & Lai [2004]), filmes (Hanjalic et al. [1999]), seriados (Li et al. [2011]), vídeos caseiros (Peng-2009) e *rushes videos* (Wang et al. [2007], Ren et al. [2007], Christel et al. [2008] e Toharia et al. [2008]). Alguns dos trabalhos classificados como genéricos são: Gong & Liu [2003], Furini et al. [2010] e Li et al. [2010].

A segunda decisão é a **determinação da duração do resumo**, que pode ser especificada no início do processo de sumarização (*A Priori*) ou no fim do processo (*A Posteriori*).

A duração do resumo determinada ***A Priori*** pode ser especificada como a duração em tempo ou uma relação sobre o comprimento l do vídeo original. Truong & Venkatesh [2007] formalizam o problema da geração do resumo dinâmico ideal como o problema de encontrar o conjunto $(E_1, ..., E_k)$, tal que o resumo produzido seja pelo menos diferente do vídeo original em relação a uma certa perspectiva p (o tipo de conteúdo a ser preservado no resumo) de sumarização.

$$K = argmin\{D(E_{i_1} \odot E_{i_2} \odot ... \odot E_{i_k}, V, p)| \sum_{i=1}^{k} |E_i| = l, E_i \subset V\}, \tag{2.1}$$

onde E_i é o i-ésimo segmento a ser incluído no resumo, $\odot$ é a operação de montagem e integração dos segmentos, D é a medida de similaridade e V o vídeo. Assim, se a perspectiva p é dada pela informatividade/cobertura, então a geração do resumo se torna um problema de otimização, onde o objetivo é encontrar a combinação que fornece a maior quantidade de informação sobre o vídeo original. Dentre os trabalhos que têm o comprimento do sumário definido a priori, podem ser citados Beran et al. [2007], Ma et al. [2002], Yamasaki et al. [2008], Lie & Hsu [2008] e Furini et al. [2010].

Por outro lado, a duração do resumo determinada ***A Posteriori***, é obtida por características implícitas do vídeo. Isso implica que o resumo gerado não é muito diferente do vídeo original em relação a perspectiva selecionada. Assim, o resumo K pode ser formalmente definido como,

$$K = arg\ mim_{E_i}\ \{\sum_{i=1}^{k} |E_i| \mid D(E_{i_1} \odot E_{i_2} \odot ... \odot E_{i_k}, V, p) < \epsilon\}, \tag{2.2}$$

onde ϵ é o limiar de duração. Por exemplo, se a perspectiva p são os eventos que ocorrem no vídeo, então a geração do resumo se torna o problema de criar uma sequência mínima

que contenha todos os eventos interessantes contidos no vídeo. Alguns exemplos de trabalhos com essa características são: Lee et al. [2003], Valdés & Martínez [2007], Furini et al. [2010] e Lu et al. [2004].

2.2 Processo de geração do resumo

Dado que as decisões iniciais foram tomadas, então é iniciado o processo de geração do sumário dinâmico. Em linhas gerais, são identificados cinco passos no processo de geração de resumos dinâmicos: segmentação, seleção, delimitação, integração multimodal e apresentação. Todavia, alguns desses passos podem ser suprimidos ou combinados. A seguir detalha-se cada um dos passos.

2.2.1 Segmentação

A segmentação divide o vídeo em suas unidades básicas (ex. tomadas). Esse é um passo essencial no processo de sumarização dinâmica. Após esse processo, as unidades do vídeo são descritas, e algumas serão selecionadas para compor o resumo.

Na literatura, as técnicas mais simples dividem o vídeo em segmentos de tamanho arbitrário, como os trabalhos de Dumont & Mérialdo [2010] e Beran et al. [2007], que dividem o vídeo em segmentos de um segundo. A maior parte dos trabalhos encontrados na literatura segmentam o vídeo através de detecção de transições entre quadros e normalmente essas transições são detectadas baseadas em similaridade visual (Pan et al. [2007], Putpuek et al. [2008]).

Algumas técnicas de segmentação levam em conta informação textual, como legendas ou o *script* do vídeo, enquanto outras, como a empregada por Taskiran et al. [2002, 2006], fazem primeiro a transcrição do áudio e em seguida, segmentam o vídeo baseado nas pausas detectadas nos diálogos do texto extraído. Para técnicas de sumarização baseadas em detecção de eventos interessantes, o processo de segmentação se resume em dividir o vídeo em sequências de eventos e não eventos.

Outras técnicas são baseadas em detecção de mudanças de movimentação (Peyrard & Bouthemy [2005]), aprendizado de máquinas (Chiu et al. [2000]), e alguns trabalhos fazem a combinação de técnicas para a segmentação do vídeo em diferentes níveis (Pan et al. [2007], Chen et al. [2008]).

2.2.2 Seleção

Após a segmentação do vídeo, a próxima etapa é a seleção dos segmentos que irão fazer parte do resumo. Essa é uma etapa muito importante, pois a técnica utilizada influenciará diretamente a qualidade do resumo, assim como o contexto e coerência. A técnica de seleção deve ser específica para o tipo de perspectiva almejada no resumo. Por exemplo, em resumos que objetivam eventos interessantes, a técnica de seleção deve escolher somente segmentos que possuam eventos de interesse.

Técnicas de eliminação de redundância inicialmente agrupam os segmentos semelhantes e de cada grupo, é escolhido um segmento representativo. Gong & Liu [2003] escolhem o segmento de maior duração para representar cada grupo, mas essa escolha também pode ser feita levando em consideração o segmento que esteja mais próximo do centróide (Pan et al. [2007], Putpuek et al. [2008]).

Alguns trabalhos definem fórmulas para a escolha dos segmentos, como em Liu et al. [2007], onde a fórmula leva em consideração a presença de objetos e eventos de áudio. Byrne et al. [2007] usaram uma abordagem que leva em consideração a quantidade de movimentação e a quantidade de faces que aparecem nos segmentos.

Outras técnicas são empregadas para a seleção dos segmentos que irão compor o resumo: modelos de atenção visual (Li et al. [2010]), eventos interessantes (Tjondronegoro et al. [2004]), quantidade de atividade/movimentação (Sasongko et al. [2008]), classificação supervisionada (Li et al. [2011]) e algoritmos bio-inspirados (Ellouze et al. [2008]).

2.2.3 Delimitação dos segmentos

Depois de selecionados, os segmentos mais importantes devem ser delimitados. Um procedimento de delimitação correto deve ser seguido para garantir que os segmentos sejam concisos, sem uma perda notável de informação. No entanto, delimitar inadequadamente um trecho de vídeo pode criar pontos de corte inadequados (por exemplo, no meio de uma fala), reduzindo a coerência global e dificultando a compreensão do telespectador. O método mais simples para delimitar um segmento é selecionar uma porção predeterminada do mesmo.

Por exemplo, Gong & Liu [2003], Lee et al. [2003] e Liu et al. [2008] definem um período de tempo para cada segmento. Dumont & Mérialdo [2010] e Beran et al. [2007] definem que todos os segmentos terão duração de um segundo. Cooper & Foote [2002] utilizam uma matriz de auto-similaridade para selecionar a parte contígua do segmento que é mais similar com o segmento completo. Ma et al. [2002] delimitam cada tomada,

identificando quadros-chave a partir de uma curva de atenção visual e utilizam apenas os segmentos em torno destes quadros-chave.

Já Pan et al. [2007] delimitam o segmento por similaridade visual entre seus quadros, e em um passo seguinte são escolhidos somente os quadros desses segmentos que possuam movimentação maior que um determinado limiar. Detyniecki & Marsala [2007] utilizam uma técnica de aceleração adaptativa, onde a ideia é apagar sucessivos quadros considerados não informativos, e exibir os quadros restantes a uma taxa constante.

Segundo Zhao et al. [2003], se trechos selecionados puderem ser classificados em diferentes classes semânticas, então sua duração pode ser reduzida proporcionalmente à preferência do usuário por essas classes. No entanto, se a perspectiva de sumarização for eventos interessantes, este passo é geralmente omitido, permitindo ao usuário uma compreensão completa dos eventos.

2.2.4 Integração multimodal e montagem do resumo

Os limites dos segmentos são muitas vezes gerados a partir de uma única fonte (áudio, imagem ou texto). O objetivo da integração é combinar as modalidades, realinhar os limites dos segmentos, e combinar todos os segmentos no resumo final. A integração multimodal é uma etapa importante pois, se feita corretamente, pode melhorar a cobertura, o contexto e a coerência do resumo dinâmico.

Os resumos baseados na integração de informações de audio e vídeo são classificados em dois tipos: sincronizado e não sincronizado. No primeiro caso, o fluxo de áudio e o fluxo visual são sincronizados de acordo com a ordem temporal do vídeo original. Neste tipo de resumo, o áudio que é ouvido corresponde diretamente à informação visual que é mostrada no vídeo, sendo muito utilizado para filmes, seriados e programas de televisão. Erol et al. [2003] e Agnihotri et al. [2004] fazem a sincronização utilizando os segmentos gerados baseados em informações visuais e de áudio.

Para vídeos de noticiários e documentários, um resumo não sincronizado pode ser útil na maximização da cobertura da informação. Resumos deste tipo também podem ser baseados em informações visuais, áudio ou texto. Por exemplo, em Smith [1997], os elementos visuais são inseridos nos segmentos baseados em áudio com base em um conjunto de regras heurísticas a respeito de movimento de objeto/câmera, faces e legendas. Para noticiários, às vezes, é melhor manter apenas o áudio do âncora. Gong & Liu [2003] integram segmentos baseados em informações visuais e de áudio através de um algoritmo de alinhamento baseado em grafo bipartido.

Para a montagem do resumo, o método mais utilizado coloca os segmentos escolhidos para compor o resumo em ordem cronológica. De fato, quase todos os trabalhos encontrados na literatura utilizam essa estratégia. A exceção é Lienhart et al. [1997], que divide os segmentos em classes, e de acordo com essas classes, a ordem cronológica pode ser quebrada. Ren et al. [2008], Beran et al. [2007] e Chen et al. [2008] utilizam transições de vídeos entre os segmentos, objetivando melhorar a coerência e entendimento do resumo. Ren et al. [2008] e Beran et al. [2007] também incorporam informações textuais e indicações visuais da posição do resumo em relação à linha do tempo do vídeo original. A avaliação desses trabalhos mostrou que a incorporação de transições e informações textuais tornou mais fácil a compreensão dos resumos (Over et al. [2007],Over et al. [2008]).

2.3 Perspectiva preservada

Para a geração de um resumo dinâmico, deve-se decidir qual a perspectiva que deve ser preservada, uma vez que diferentes perspectivas criam diferentes formas de resumos. De acordo com a Figura 2.1, são identificados três tipo de perspectivas: informação/cobertura, eventos importantes/interessantes e contexto de pesquisa/personalização.

2.3.1 Cobertura da informação

Nesse tipo de perspectiva, o objetivo é fornecer uma impressão do vídeo por completo, sem afetar muito a compreensão do usuário. Geralmente, são empregadas técnicas de eliminação de redundância para a geração desse tipo de resumo. São mais aplicados a vídeos baseados em informação, como vídeos instrucionais e noticiários, onde os usuários estão mais interessados na informação do vídeo como um todo. Os seguintes trabalhos objetivam a preservação da informação e cobertura do conteúdo: Gong & Liu [2003], Ren et al. [2007], Pan et al. [2007] e Putpuek et al. [2008].

2.3.2 Eventos importantes/interessantes

O objetivo desse tipo de resumo é preservar eventos interessantes ou importantes que aparecem no vídeo. Eventos interessantes/importantes são definidos de acordo com o contexto do tipo de vídeo que está sendo trabalhado. São mais aplicados a vídeos esportivos, onde o conceito "interessante" pode ser definido, modelado e extraído. Com-

parado a outros tipos de vídeos como notícias e filmes, vídeos de esportes têm estrutura de conteúdo e regras de domínio bem definidas.

Em vídeos esportivos, um segmento de vídeo é definido como importante/interessante se:

- prende a atenção do espectador, ou seja, apresenta grande movimentação ou cores chamativas (Ma et al. [2002]);
- possui padrões específicos, raros ou incomuns (Xiong et al. [2003]);
- possui padrões específicos de visão, vistos com maior frequência pelo usuário (Yu et al. [2003]);
- é um evento com um rótulo semântico específico como por exemplo, a expulsão de um jogador em vídeos de futebol (Assfalg et al. [2003]);
- possui atributos temporais e espaciais específicos, como por exemplo, que ocorrem próximos a uma das traves em um jogo futebol (Wang et al. [2004]);
- provoca determinadas reações da platéia, por exemplo, aplauso ou vaias (Xiong et al. [2003]);
- provoca determinadas reações do narrador/comentarista, por exemplo, grito de gol (Coldefy & Bouthemy [2004]);
- provoca determinadas reações do produtor de conteúdo, por exemplo, mudança na taxa de transições, na legenda que informa placar ou *replays* (Tjondronegoro et al. [2004]).

Um dos grandes problemas desse tipo de perspectiva é definir os limites dos eventos detectados, garantindo concisão ao resumo e ao mesmo tempo preservando o contexto do evento.

2.3.3 Contexto de consulta e personalização

Resumos desse tipo são gerados com base em especificações fornecidas pelos usuários, seja em uma consulta ou personalização. Eles solucionam o problema de decidir quais os eventos de destaque devem ser selecionados para compor o resumo, uma vez que são mantidos os segmentos que forem mais relevantes para uma consulta ou necessidades dos usuários.

Geralmente os usuários personalizam os resumos através de atribuições de pesos sobre um conjunto de características (incluindo tipo de evento) associado a segmentos candidatos. Por exemplo, as características relevantes em Lu et al. [2004] são (faces, fala, zoom de câmera, legendas) para vídeo de notícias e (disparo de revolver/explosão, voz alta, face, chamas) para filmes. Para resumos de vídeos caseiros, em Zhao et al. [2003], pesos são requeridos para eventos temporais exclusivos (fala, riso, música, aplausos, gritos, movimento). Para filmes, Li et al. [2004] propõem um conjunto de seis características (taxa musical, taxa de fala, volume de som, elenco atual, nível de ação e tema do tópico) extraído de qualquer evento de história (diálogo entre duas pessoas, diálogo com várias pessoas, e cena híbrida). Em vídeos esportivos, os resumos podem ser personalizados, por exemplo, de acordo com o evento preferido, jogador preferido ou equipe preferida. Por outro lado, Agnihotri et al. [2005] ao invés de solicitar ao usuário especificações, tenta produzir resumos que são adaptáveis à personalidade do usuário, por exemplo, masculino/feminino, introvertido/extrovertido.

2.4 Mecanismo de geração do resumo

São identificados três tipos fundamentais de mecanismo para a geração de resumos (Figura 2.1): eliminação de redundância, detecção de eventos e formulação de curva de resumo. Normalmente a perspectiva almejada no resumo define o tipo de mecanismo utilizado. Por exemplo, resumos que objetivam a detecção de eventos interessantes não irão obter bons resultados relativos a preservação do aspecto de cobertura do vídeo.

2.4.1 Eliminação de redundância

Esse mecanismo elimina segmentos ou partes de segmentos que contenham informação similar. Ele pode ser realizado na etapa de seleção ou de delimitação dos segmentos.

Furini et al. [2010] eliminam redundância através do algoritmo de agrupamento FPF (*Farthest Point-First*), onde os segmentos são representados por histogramas de cor no espaço HSV. Um segmento de cada grupo é escolhido para formar o resumo. Pan et al. [2007] eliminam a redundância em duas etapas, primeiro é aplicado um algoritmo de agrupamento hierárquico aglomerativo, onde para cada grupo é escolhido um segmento representativo. Em uma etapa posterior é calculada a movimentação entre os quadros de um segmento e somente os que estiverem acima de um limiar de movimentação são escolhidos para compor o resumo final. Laganière et al. [2008] geram uma matriz espaço-temporal de pontos de interesse, que é utilizada para eliminar segmentos similares.

Na faixa de áudio, a eliminação de redundância pode ser feita removendo trechos que possuam ruído ou silêncio. Como a remoção de redundância leva em consideração somente informação proveniente de um único canal (visual, áudio ou texto), um problema crucial surge: como combinar a detecção em mais de um canal, uma vez que segmentos identificados como redundantes em um canal podem conter informação relevante em outro.

2.4.2 Detecção de eventos interessantes

O objetivo da detecção de eventos é identificar e localizar padrões espaço-temporais específicos, uma pessoa acenando sua mão. Ela engloba não somente a detecção do evento, mas também sua delimitação.

Hannon et al. [2011] usam uma técnica que combina termos de pesquisa (exemplo, cartão vermelho, gol, *penalty*) fornecidos por usuários e informações extraídas da rede social *Twitter* para a detecção de eventos em vídeos de futebol.

Dagtas & Abdel-Mottaleb [2004] detectam eventos de destaque em vídeos de esportes através de dois métodos: localização de co-ocorrência de palavras-chave com áudio de grande energia, e detecção de transições entre sequências de quadros de grandes áreas com gramado (ação do jogo) para sequências de quadros que não incluem grandes áreas gramadas (*closes*).

Em Peyrard & Bouthemy [2005], os eventos esportivos são detectados levando em conta apenas características de movimento. Inicialmente, o vídeo é segmentado baseado em suas características de movimentação. Então, para cada segmento, é adotada uma representação estatística do conteúdo de movimentação para classificar em classes pré-determinadas (por exemplo, salto com vara, segmentos de entrevista, grandes vistas do estádio, etc, para um vídeo de atletismo), de acordo com o princípio da máxima verossimilhança. Assfalg et al. [2003] detectam eventos de futebol a partir de modelos de lógica temporal. Cada modelo representa a relação entre as zonas do campo de jogo, movimento da bola, e as posições dos jogadores para um evento específico.

Em vez de detectar eventos de jogos específicos, os segmentos de interesse podem ser identificados através da detecção de eventos de áudio, visuais ou cinematográficos, que seriam gerados a partir de eventos importantes do jogo (Tjondronegoro et al. [2004]).

2.4.3 Formulação de curva de resumo

Nesse processo, a seleção é feita associando os segmentos a uma curva de perspectiva. Segmentos que possuírem valores acima da curva são então selecionados para compor o resumo. Os segmentos são representados por características que definem pontuações referentes à perspectiva objetivada. Por exemplo, se um resumo é baseado em eventos interessantes, as características devem refletir o quanto cada segmento é interessante.

A grande questão desse método é como calcular as medidas de interesse em relação as unidades do vídeo. Para vídeos de esporte, uma maneira de calcular medidas de interesse seria associar a medida de interesse ao tipo de evento detectado, ou a reações detectadas na platéia, como o volume ou a duração de momentos em que a platéia emite sons, tais como, aplausos ou vaias (Lu et al. [2005]). Caso as unidades do vídeo sejam baseadas em texto, técnicas de sumarização de texto podem ser utilizadas. Para calcular medidas de interesse sobre unidades de vídeos genéricos, um conjunto de características é definido sobre as unidades do vídeo, e utilizando modelos genéricos é calculada uma medida de interesse onde cada característica tem uma parcela de contribuição (Song & Wang [2009]).

A Figura 2.2 mostra o processo mais simples para a geração de resumos baseados em curva de perspectiva. Nesse processo é definido um limiar sobre a curva de perspectiva, e os segmentos com valores de perspectiva acima desse limiar são então selecionados para compor o resumo.

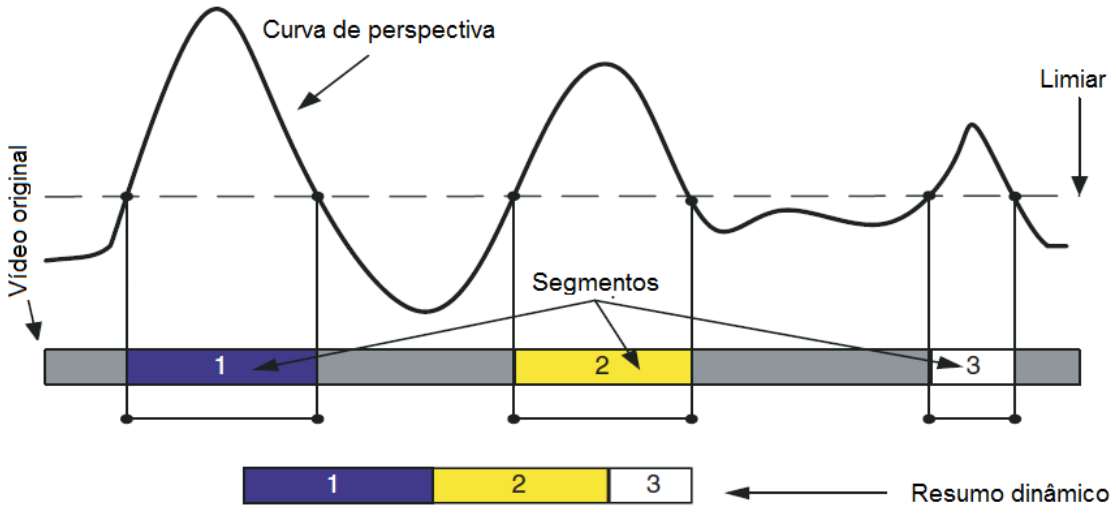

Figura 2.2. Geração de resumo dinâmica através de curva de perspectiva (Truong & Venkatesh [2007]).

Se a unidade básica for um quadro, um segmento de tamanho fixo ou uma to-

mada, a geração do resumo como demonstrada na Figura 2.2 não assegura coerência e cobertura balanceada do conteúdo, uma vez que o resumo pode conter segmentos muito curtos ou similares. Lu et al. [2005] tentam resolver esse problema descobrindo um conjunto mínimo de segmentos que maximiza o valor total da medida de importância através de um método guloso. Já em Mei et al. [2005] a curva é formada para vídeos caseiros estimando a qualidade em relação a um conjunto de métricas, mas ao invés de definir um limiar para a curva, a cobertura é assegurada através da seleção uniforme de subtomadas ao longo do vídeo, modelando a seleção de segmentos como um problema de otimização.

2.5 Características

A escolha das características utilizadas no processo de summarização é uma etapa muito importante, pois a representação dos dados influencia no tipo de mecanismo utilizado para a sumarização, bem como, na qualidade do resumo produzido. A seguir, são expostas algumas categorias encontradas na literatura:

- **Visuais** - Características visuais são usadas geralmente para medir a similaridade entre quadros/segmentos, etapa essa que é essencial para os métodos baseados em eliminação de redundância. Abordagens como as de Pan et al. [2007], Le & Satoh [2007] e Furini et al. [2010] utilizam histogramas locais de cor para realizar a detecção de tomadas e agrupamento dos segmentos do vídeo. Já Gong & Liu [2003]; Lu et al. [2004]; Ma et al. [2002]; Lee et al. [2003] calculam características baseadas em contraste para criar um modelo do nível de atenção humana para uma determinada imagem. Em vídeos esportivos, elementos visuais como cor dominante, bordas, textura e sua posição espacial desempenham um papel muito importante na identificação de eventos de camera (visão geral da cena, *close* em jogadores, cenas próximas ao gol, etc.), os quais podem ser utilizados para estimar a ocorrência de eventos interessantes (Chang et al. [2002]; Assfalg et al. [2003]; Tjondronegoro et al. [2003]).

- **Texto** - Texto é uma importante característica, uma vez que conceitos semânticos podem ser extraídos mais facilmente quando comparado a características visuais e de áudio. O texto pode ser obtido de diversas formas: a) diretamente do fluxo de vídeo (Shao et al. [2006]); b) legendas (Miyauchi-2003); c) transcrição de áudio (Gong & Liu [2003]); d)fontes externas (Hannon et al. [2011]). Normalmente, a informação textual é utilizada através da identificação de palavras-chave, que são

bons indicadores para a identificação de eventos interessantes em vídeos esportivos. Porém, a maioria das fontes de vídeos atuais não disponibilizam informações textuais. Para lidar com essa dificuldade, Li et al. [2011] propõem uma abordagem baseada em transferência de conhecimento, onde é utilizada uma base de treinamento que possui informação textual, as informações textuais são usadas para treinar um modelo de sumarização. Esse modelo leva em conta somente características visuais para sumarização de um novo vídeo.

- **Áudio** - São utilizadas características de baixo-nível extraídas da banda do diálogo, que podem ser associadas ao nível de interesse do usuário em determinado segmento do vídeo. Também são utilizadas para construir modelos para a detecção de efeitos de sons genéricos, associados ao entusiasmo em várias categorias de vídeos (torcida, aplausos, gargalhadas, etc.) (Coldefy & Bouthemy [2004]; Wang et al. [2004]). Alternativamente, eventos de som associados a um tipo específico de vídeo podem detectar segmentos com grande probabilidade de serem eventos interessantes (Tjondronegoro et al. [2004]).

- **Visuais dinâmicas** - Normalmente, são indicativos de interesse/importância para o usuário. Pan et al. [2007] argumentam que quanto mais dinamismo há em uma cena, maior é a quantidade de informação fornecida. Essas características podem ser usadas para a detecção de ações em cenas (Hanjalic et al. [1999]), eventos de interesse em vídeo esportivos (Peyrard & Bouthemy [2005]). Também são utilizadas para direcionar a atenção do usuário para um determinado momento da sequência do vídeo, como modelado em métodos que utilizam formulação da curva do resumo (Ma et al. [2002]; Pan et al. [2007]; Le & Satoh [2007]).

- **Movimentos de câmera** - São utilizados principalmente em vídeo esportivos e *rushes videos*, pois a posição da camera nesses tipos de vídeos está fortemente relacionada a ocorrência de eventos importantes (Assfalg et al. [2003]; Coldefy & Bouthemy [2004]; Wang et al. [2007]).

- **Semânticas** - Essa categoria se refere ao uso de características semânticas no processo de sumarização. Características semânticas podem ser indicativos de importantes eventos em sequência de vídeos. Em Ren et al. [2007] a informação de ocorrência de faces é utilizada na fórmula para a escolha de segmentos mais representativos do vídeo. Para vídeos onde dados de produção são disponíveis, por exemplo *rushes videos*, informações adicionais podem ser incorporadas (marcação de inicio e fim de gravações de cenas, presença de elenco, etc.) (Erol et al. [2003]).

2.6 Métodos de avaliação

A área de sumarização de vídeos ainda não possui um modelo de avaliação consolidado, e as metodologias propostas em sua maioria possuem graves deficiências. Isso faz com que cada trabalho crie sua própria metodologia de avaliação, o que muitas vezes não permite a comparação com outros trabalhos da literatura. Isso ocorre principalmente devido a falta de bases de dados de domínio público, que possuam anotações do que deve conter em um resumo (*ground truth*).

Outra dificuldade encontrada é a reprodução dos trabalhos da literatura, pois geralmente os sistemas existentes não são disponibilizados para avaliações comparativas, ou complicados de se utilizar por trabalharem com dados em formatos e configurações diferentes.

As principais formas de avaliação de resumos dinâmicos são agrupadas em duas categorias: métricas objetivas e avaliação através de usuários (Truong & Venkatesh [2007]).

Para a utilização de **métricas objetivas** na avaliação de resumos dinâmicos, necessita-se que o resumo tenha sido gerado sobre uma base de dados que possua *ground truth*. Assim, métricas comuns de recuperação de informação podem ser aplicadas (ex. precisão e revocação). Em geral, esse tipo de avaliação só é empregada para trabalhos baseados em detecção de eventos interessantes, principalmente em vídeos esportivos (Chang et al. [2002]; Xiong et al. [2003]; Song & Wang [2009]), uma vez que para esse tipo de perspectiva é relativamente simples a criação de um *ground truth*. Porém para eventos genéricos, o *ground truth* pode ser muito subjetivo (Tjondronegoro et al. [2004]).

Uma alternativa empregada por alguns trabalhos é a utilização de resumos criados manualmente por usuários para alcançar uma avaliação mais objetiva (Reiko et al. [2003]; Takahashi et al. [2005]). Por exemplo, He et al. [1999] utilizam resumos de vídeos de palestras feitos pelos próprios palestrantes. Reiko et al. [2003] utilizam resumos gerados pelo produtor que acompanha um programa de culinária. Takahashi et al. [2005] usa os melhores momentos transmitidos por emissoras de TV.

Já **métricas mais subjetivas** envolvem usuários independentes avaliando a qualidade dos resumos, e representam provavelmente a mais útil e realista forma de avaliação. Os usuários são questionados a diretamente quantificar a qualidade do resumo, bem como informar o grau de satisfação ao assistir o resumo, ou ainda reportar o quanto o sumário o ajudou a realizar tarefas como visualização, pesquisa e identificação de conteúdo em bases de vídeos (Sundaram et al. [2001]; Ma et al. [2002]; Li et al. [2004]).

Um grande avanço ocorreu no TRECVID 2007, onde foi definida uma metodologia de avaliação por usuário e disponibilizada uma base anotada. A qualidade dos resumos foi avaliada por medidas subjetivas, tais como: fração de segmentos importantes do vídeo completo incluídos, facilidade em encontrar o conteúdo desejado e a quantidade de conteúdo redundante presente no resumo. No total, 53 trabalhos foram avaliados nas edições de 2007 e 2008 (Over et al. [2007, 2008]).

Uma outra alternativa foi proposta por Dumont & Mérialdo [2010]. Uma automatização da avaliação manual do TRECVID foi implementada utilizando técnicas de aprendizado de máquina para treinar avaliadores automáticos. Para isso, os autores adicionaram manualmente informações ao *ground truth* fornecido pelo TRECVID, incluindo limites precisos de tempo das ocorrências dos tópicos presentes no *ground truth*. Os resultados mostraram alta correlação entre a avaliação automática e a avaliação manual realizada no TRECVID. Uma discussão da metodologia do TRECVID 2007, também utilizada nesse trabalho, é feita na Seção 4.3.

Capítulo 3

Uma nova metodologia para sumarização de *rushes videos*

Neste capítulo, é descrita a metodologia criada para gerar resumos dinâmicos de *rushes videos*. Como mencionado anteriormente, o conteúdo de vídeos *rushes videos* é muito específico. *Rushes videos* contêm uma grande quantidade de repetições, muitas vezes com pequenas variações. Eles também podem conter longos segmentos em que a câmera é fixa em uma determinada cena ou com pouco movimento, e tomadas reutilizáveis de pessoas, objetos, eventos, locais, que às vezes são usados para preencher lacunas durante a edição final.

Rushes videos também contêm muitas sequências irrelevantes de quadros, chamados *junk shots*, que são por exemplo, quadros padrões de teste, quadros de cor uniforme, sequências de claquetes e etc.

Embora muitas técnicas tenham sido propostas para processar automaticamente o conteúdo de vídeos em geral, a estrutura específica dos *rushes videos* requer uma adaptação destas técnicas e, às vezes, o desenvolvimento de novas abordagens para uma análise eficiente (Dumont & Mérialdo [2010]).

A Figura 3.2 apresenta o esquema geral da metodologia proposta para sumarização dinâmica de vídeos. Inicialmente, o vídeo é segmentado em suas unidades básicas. No passo seguinte, são eliminados os segmentos que não possuam informação relevante para compor o resumo. Em seguida, os segmentos são descritos utilizando três descritores, sendo um descritor espaço-temporal (STIP (Laptev [2005])) e dois descritores espaciais (SIFT (Lowe [2004]) e HueSIFT (van de Sande et al. [2010])), associados a uma estratégia baseada em histogramas de palavras visuais (BoVFs). A Figura 3.1 mostra em vermelho as contribuições feitas no presente trabalho e em azul os passos utilizados que são comuns à literatura pesquisada.

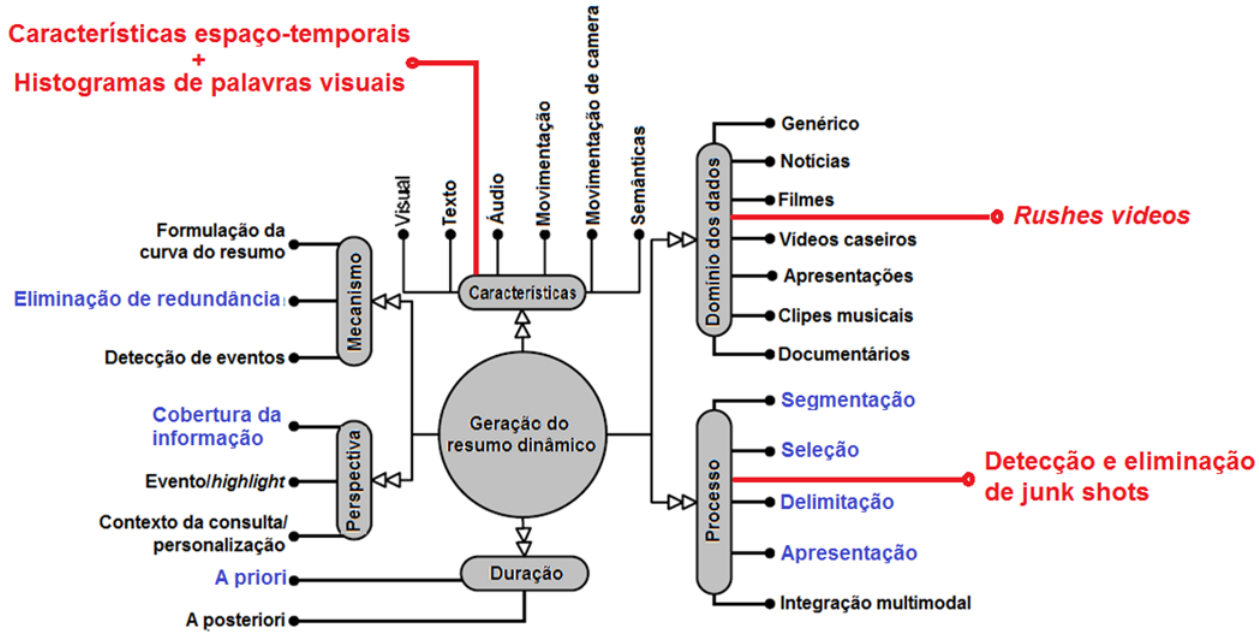

Figura 3.1. Contribuições da abordagem proposta (adaptado de (Truong & Venkatesh [2007])).

Após a descrição, cada segmento é representado por um histograma de palavras visuais, os quais são agrupados e de cada grupo é escolhido um segmento-chave. O agrupamento objetiva remover redundância entre os segmentos do vídeo, e como sequências repetitivas são suscetíveis de serem tomadas diferentes da mesma cena, agrupando-as, pode-se identificar as várias cenas que ocorrem no vídeo. Nó próximo passo são escolhidos os segmentos mais significativos que irão compor os resumos associados à cada descritor. Por fim, os segmentos dos três resumos gerados são integrados e são escolhidos os segmentos mais significativos da integração. Estes segmentos são concatenados em ordem cronológica, formando assim o resumo dinâmico do vídeo. Cada um destes passos é detalhado a seguir.

3.1 Segmentação de vídeo

Um arquivo de vídeo é organizado em uma estrutura hierárquica. O nível mais baixo consiste em um conjunto de quadros. No próximo nível, os quadros são agrupados formando tomadas, onde cada tomada é composta por uma sequência contínua de quadros capturados por uma única câmera. Tomadas relacionadas a um local comum ou evento são agrupadas em cenas. Um grupo de cenas forma o vídeo.

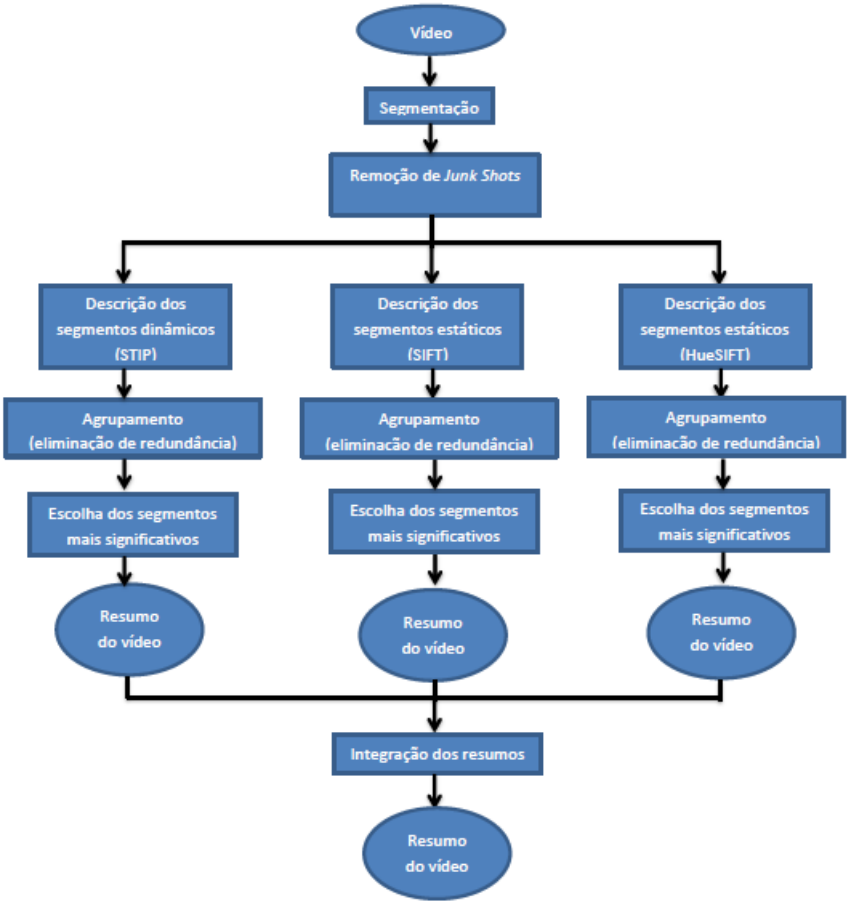

Figura 3.2. Arquitetura da abordagem proposta.

Uma vez em que o vídeo é dividido em suas unidades significativas e gerenciáveis, pode-se iniciar o processo de caracterização desses componentes individuais. Normalmente, a segmentação do vídeo é o primeiro passo em direção à indexação e anotação automática de vídeos.

Segmentação do vídeo refere-se a particionar o vídeo de maneira espacial, temporal ou espaço-temporal em regiões que são homogêneas em algum espaço de características (Bovik [2009]).

Nesse trabalho, a segmentação é feita em quatro fases, como ilustrado na Figura 3.3:(i)detecção de tomadas, (ii) detecção de subtomadas, (iii) detecção de *junk shots* e (iv) limitação do tamanho máximo dos segmentos.

Para realizar a segmentação dos vídeos, uma etapa de delimitação dos segmentos foi adicionada ao método proposto por Pan et al. [2007]. Esse método foi escolhido por possuir um baixo custo computacional, além de alcançar bons resultados em relação à detecção de tomadas e *junk shots* (Over et al. [2007]). Ele se beneficia do uso de histo-

gramas locais de cor e vetores de movimentação, que são características eficientemente computáveis.

Histogramas locais de cores são utilizados para segmentação em tomadas, subtomadas e detecção de *junk shots*. Vetores de movimento são utilizados para a segmentação em tomadas. Ao fim do processo uma etapa adicional de limitação do tamanho máximo dos segmentos foi incorporada ao método, devido a análise das características dos vídeos trabalhados.

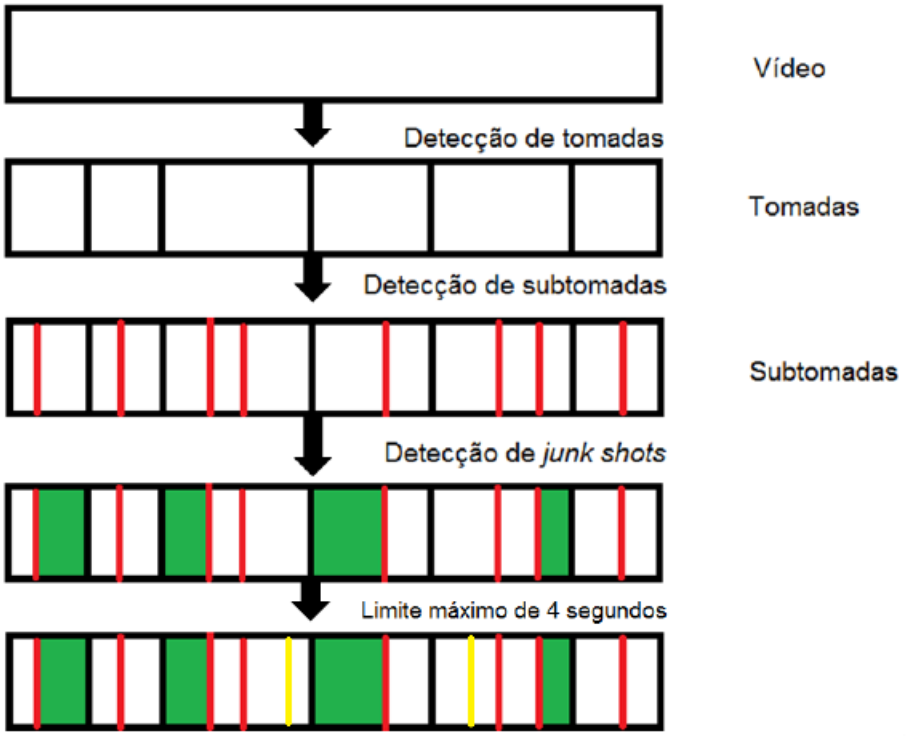

Figura 3.3. Passos da segmentação do vídeo (adaptado de Pan et al. [2007]).

3.1.1 Detecção de tomadas

Visto que uma tomada é capturada por uma única câmera, quadros adjacentes de uma mesma tomada devem exibir forte continuidade temporal. Descontinuidades ocorrem em transições de tomadas onde o conteúdo muda. Existem dois tipos básicos de transições de tomadas: abrupta e gradual. Uma transição abrupta ocorre em um único quadro entre duas tomadas e são decorrentes da junção de duas tomadas, onde nenhum quadro é criado ou modificado durante a transição. Por outro lado, a transição gradual combina duas tomadas pelo *fade-in*/*fade-out*, dissolver ou outros efeitos cinematográficos (Yuan et al. [2007]).

Transições graduais são mais difíceis de serem detectadas do que cortes, especialmente difíceis de detectar dissolver entre as sequências envolvendo movimento intenso. Porém, como *rushes* são materiais não editados, geralmente possuem somente transições abruptas. Deste modo abordagens simples conseguem obter bons resultados (Putpuek et al. [2008]). O método descrito a seguir detecta somente transições abruptas.

O método proposto por Pan et al. [2007] e adaptado para esse trabalho calcula histogramas locais de cor após dividir um quadro em blocos 4x4 usando um código de seis bits para representar as cores no espaço RGB, sendo dois bits para cada canal. H_t^k é utilizado para representar o histograma local de cor para o bloco k do quadro t, onde $k = 0...15$. H_t^k[i] representa o valor do i-ésimo *bin* do histograma, onde $i = 0...63$. Os autores reportam que foram testados outros espaços de cores, mas os resultados alcançados não foram melhores que os obtidos com o espaço RGB.

Para calcular a distância entre dois histogramas foi utilizada a distância χ^2. A distância D_{χ^2}(H,G) entre dois histogramas H e G é definida como:

$$D_{\chi^2}(H,G)\sum_{i=0}^{63} = \begin{cases} \frac{(H[i],G[i])^2}{max(H[i],G[i])} & se\ max(H[i],G[i]) > 0 \\ 0\ \ caso\ contrário \end{cases}$$

A diferença χ_k^2 entre os blocos k dos quadros t e t+1 é definida como:

$$\chi_k^2 = D_{\chi^2}(H_t^k, H_t^k + 1) \tag{3.1}$$

Para cada par de quadros t e t+1 são calculadas as diferenças entre os blocos. Então, os 16 valores são ordenados, e caso a soma dos oito valores centrais exceda um limiar predefinido (ϵ_{cut}), i.e., $\sum_{k'=4}^{11} \chi_{k'}^2 > \epsilon_{cut}$), então diz-se que há uma transição entre os quadros t e t+1.

Lupatini et al. [1998] recomendam o uso dos 8 valores iniciais de χ^2 para evitar alarmes falsos devido a grande movimentação de objetos entre os quadros. Grandes movimentações de objetos causam grandes diferenças na distribuição de cores dos blocos afetados. Porém, uma vez que o objeto ocupa menos que a metade do quadro e a câmera não muda drasticamente, os oito valores iniciais de χ^2 conterão somente informações referentes ao fundo do quadro, o que deve continuar similar em uma tomada.

Porém Pan et al. [2007] argumentam que em *rushes* utilizar os 8 valores iniciais de χ^2 resulta em baixo valor de revocação na detecção de cortes. Isso porque *rushes* contém várias tomadas regravadas com conteúdo de cena similar. Assim as transições entre essas tomadas não podem ser detectadas quando utilizada a parte inferior dos valores de χ^2. Os autores verificaram os resultados alcançados pelas três porções (iniciais, centrais e finais) dos 16 valores, e chegaram a conclusão que usar os oito valores centrais (Figura

3.4) seria a melhor escolha, pois foi a que mostrou melhor compromisso entre precisão e revocação.

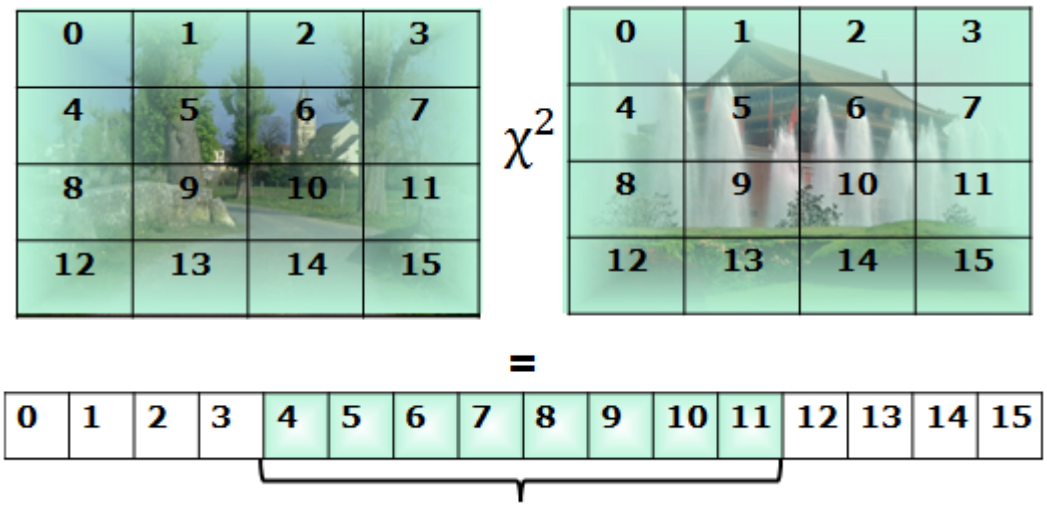

Figura 3.4. Exemplo do cálculo dos valores de χ^2.

Contudo, ao utilizar a parte central, a chance de ocorrência de falsos alarmes aumenta quando há grande movimentação de objetos e consequentemente os valores de precisão são menores do que os obtidos ao utilizar a parte inicial. Para solucionar esse problema, foi adicionado um teste que verifica se a detecção do corte ocorreu por grande movimentação de um objeto.

Além disso, são calculados vetores de movimentação entre os quadros onde foi detectado o corte. Em seguida, é feito o somatório das magnitudes de cada vetor, formando assim uma única medida de movimentação M_t. Caso M_t seja maior que um limiar de movimentação (ϵ_{motion}), diz-se que ocorreu um alarme falso e essa detecção é descartada. Com a adição dessa etapa de teste, foram obtidos melhores resultados de precisão e revocação. Para o cálculo dos vetores de movimentação foi utilizado o método de cálculo de fluxo óptico Lucas-Kanade (Lucas & Kanade [1981]), disponível na biblioteca OpenCV (Bradski [2000]).

3.1.2 Detecção de subtomadas

Em *rushes* uma tomada representa uma gravação de uma cena. No entanto, tomadas representando uma mesma cena não possuem necessariamente a mesma duração, pois cenas podem ser interrompidas prematuramente devido a erros de gravação. Trechos de vídeos com durações diferentes tendem a ser mais difíceis de serem comparados, pois a distribuição de cores nos histogramas locais de cor pode variar muito. Assim para

tentar solucionar esse problema e tornar a comparação mais confiável, as tomadas foram divididas em unidades menores denominadas subtomadas. Subtomadas são segmentos de vídeo pertencentes a uma tomada, que apresentam grande similaridade visual entre seus quadros.

Para segmentar as tomadas em subtomadas, inicialmente é escolhido o primeiro quadro da tomada como um quadro base, b. Em seguida, esse quadro é comparado sequencialmente até que haja algum quadro suficientemente diferente, c. Os quadros de b a c - 1 formam então uma subtomada e o quadro c é escolhido como o novo quadro base. Este processo é repetido até que todos os quadros da tomada sejam processados. Uma subtomada é detectada se:

$$\sum_{k'=0}^{7} D_{\chi^2}(H_b^{k'}, H_c^{k'}) > \epsilon_{subshot} \tag{3.2}$$

Na Equação, 3.2 $\epsilon_{subshot}$ representa o limiar de detecção de subtomadas e a parte inicial dos valores de χ^2 é escolhida, pois o objetivo é que a subtomada tenha grande similaridade visual evitando influência da movimentação de objetos. Uma vez que todos os quadros tenham conteúdo visual similar, o histograma local médio H_s^k é usado como uma característica que representa a subtomada s.

$$H_s^k[i] = \frac{1}{|s|} \sum_{t \epsilon s} H_t^k[i] \tag{3.3}$$

A Figura 3.5 mostra um exemplo de cálculo para a detecção de subtomada.

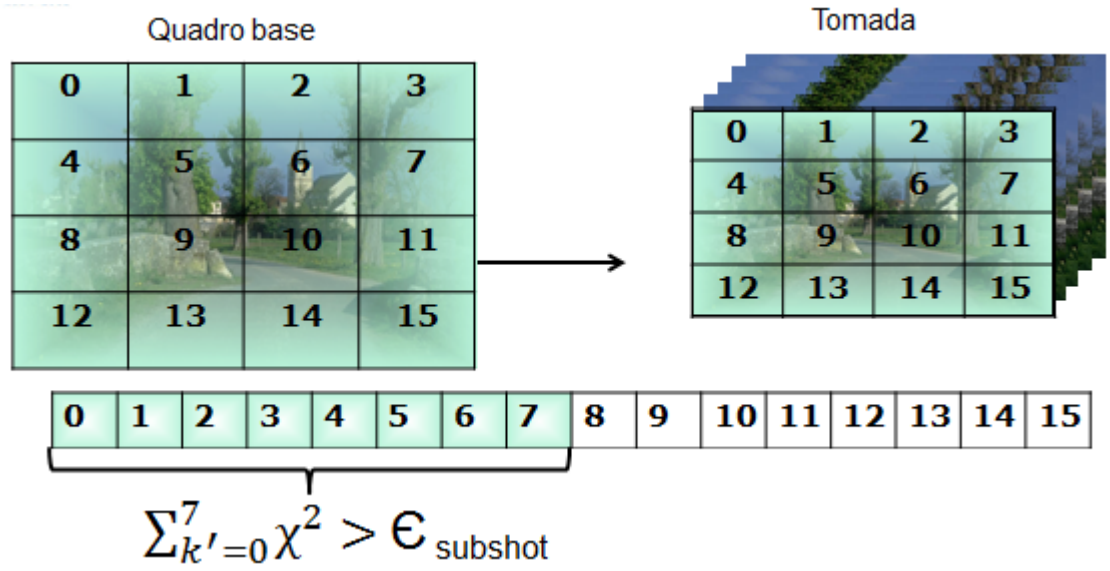

Figura 3.5. Exemplo de cálculo para a detecção de subtomada.

3.1.3 Remoção de *junk shots*

Por se tratar de material não editado, existem trechos de vídeo que são utilizados única e exclusivamente para marcação entre as gravações de cenas. Esses trechos devem ser removidos, pois não fornecem informação relevante para compor o resumo. Esses trechos são tomadas muito curtas, que possuam barras verticais de cor, claquetes ou de cor única.

A detecção de subtomadas muito curtas é bem simples, pois as subtomadas com duração inferior a um segundo são simplesmente descartadas. Para a detecção dos outros tipos de segmentos indesejados é utilizado o histograma médio, definidos na Seção 3.1.2 .

Subtomadas de cor única são detectadas se houver alguma cor dominante no seu histograma global H_s, onde $H_s[i] = \frac{1}{16}\sum_{k=0}^{15} H_s^k[i]$. Então, se max$H_s$[i]>$\epsilon_{single}$, a subtomada é classificada como de cor única.

Em barras de cor vertical, os blocos pertencentes a uma mesma coluna devem ser similares. Assim, são feitas 12 comparações entre quaisquer dois blocos vizinhos em cada coluna, $D_{\chi^2}(H_s^k, H_s^{k+4})$, com k variando de 0 a 11. Se pelo menos 10 dessas diferenças forem menores que um determinado limiar ($\epsilon_{colorbar}$), então uma subtomada com barras de cor é detectada e eliminada.

Na remoção de *junk shots* a detecção de subtomadas que possuam claquetes é a que requer um processo mais complexo, pois as claquetes podem variar de sua forma e cor.

Os trabalhos encontrados na literatura utilizam, em sua maioria, métodos de alto custo computacional e que devido a variedade na forma das claquetes não conseguem detectar todos os tipos presentes nos vídeos.

Sendo assim, decidiu-se remover as claquetes manualmente. Esse processo foi facilitado, pois após a segmentação e remoção de segmentos indesejados todas as subtomadas geradas possuem grande similaridade visual. Assim, os trechos que possuem claquetes ficam concentrados em algumas subtomadas que são facilmente identificadas na pré-visualização do conteúdo do segmento, evitando que todas as subtomadas necessitem ser assistidas por completo.

Para garantir uma maior corretude no processo, são verificadas as subtomadas anterior e a posterior aos trechos onde foram identificadas as claquetes. As subtomadas que possuem claquetes são então eliminadas.

3.1.4 Limitação do tamanho máximo dos segmentos

Após a segmentação em subtomadas, todos os componentes do vídeo possuem grande similaridade visual entre seus quadros. Porém, isso ainda pode gerar alguns problemas em relação ao tamanho dos componentes dos vídeos. Como alguns vídeos possuem grande similaridade visual entre todos os seus quadros, a segmentação em subtomadas pode gerar segmentos muito grandes, ocasionando problemas na formação do resumo, uma vez que segmentos importantes podem não ser escolhidos para compor o resumo por ultrapassarem um tamanho máximo predefinido para o resumo.

Para minimizar esse problema, uma etapa foi adicionada ao método proposto por Pan et al. [2007]. Definiu-se que as subtomadas teriam uma duração máxima. Baseado na segmentação em subtomadas feitas por Liu et al. [2008] e por observações feitas nas anotações de cada vídeo, foi definido que as subtomadas teriam duração máxima de quatro segundos. Essa definição permite que a maioria dos eventos contidos nas anotações dos vídeos sejam vistos por completo, facilitando assim a visualização e consequente avaliação feita por usuários.

3.2 Descrição das unidades básicas do vídeo

No método proposto, as unidades básicas do vídeo são descritas por dois descritores espaciais e um descritor espaço-temporal. Esses descritores detectam e descrevem pontos de interesse, que são pontos específicos (ou regiões) em uma imagem ou vídeo que apresentam significante variação de intensidade em mais de uma direção. Sendo amplamente utilizados em visão computacional para tarefas de rastreamento, comparação e reconhecimento (Yan et al. [2004]; Laganière et al. [2008]).

Para a representação dos segmentos definidos anteriormente é utilizada uma estratégia baseada em histogramas de palavras visuais espaço-temporais (*bags-of-visual-features*) (BoVF). BoVF é uma técnica livre de modelo, ou seja, onde nenhum conhecimento prévio é dado. São inspirados em técnicas tradicionais de recuperação de informação textual, em que um documento é visto apenas como uma coleção de palavras, e a ordem de ocorrência dessas palavras não é considerada. Histogramas de palavras visuais são uma forma robusta de representação de imagens/vídeos, onde cada imagem/vídeo é vista como um conjunto de regiões, onde a informação espacial não é levada em consideração, mas somente a aparência da região (deCampos et al. [2011]).

No caso de imagens, muitas abordagens utilizam pontos de interesse espaciais similarmente às famílias de palavras, os pontos recebem o nome de palavras visuais. Os pontos pertencentes a uma mesma imagem são agrupados baseados em uma medida

de similaridade, mostrando bons resultados para reconhecimento de objetos devido a robustez à luminosidade, escala, posição e oclusão (Tuytelaars & Mikolajczyk [2008]).

Com o sucesso obtido para o reconhecimento de objetos, abordagens estenderam seu uso para o reconhecimento de ações. Muitas das técnicas de reconhecimento de ações assumem que o vídeo é um volume espaço-temporal, selecionando assim pontos de interesse 3D. Entretanto, há abordagens alternativas que utilizam descritores 2D aplicados ao domínio espaço-temporal dos vídeos. Essa abordagem alcança resultados similares às abordagens em 3D, porém com a desvantagem de possuírem um custo computacional mais elevado (Jiang et al. [2007]).

BoVF fornece uma representação mais informativa do vídeo em termos de padrões de baixo nível, espera-se que as características que compõem o histograma sejam padrões de fato representativos para descrever o conteúdo do vídeo. BoVFs tentam diminuir o gap semântico entre as características de baixo-nível e o conteúdo do vídeo. A Figura 3.6 mostra os passos para o cálculo dos histogramas de palavras visuais.

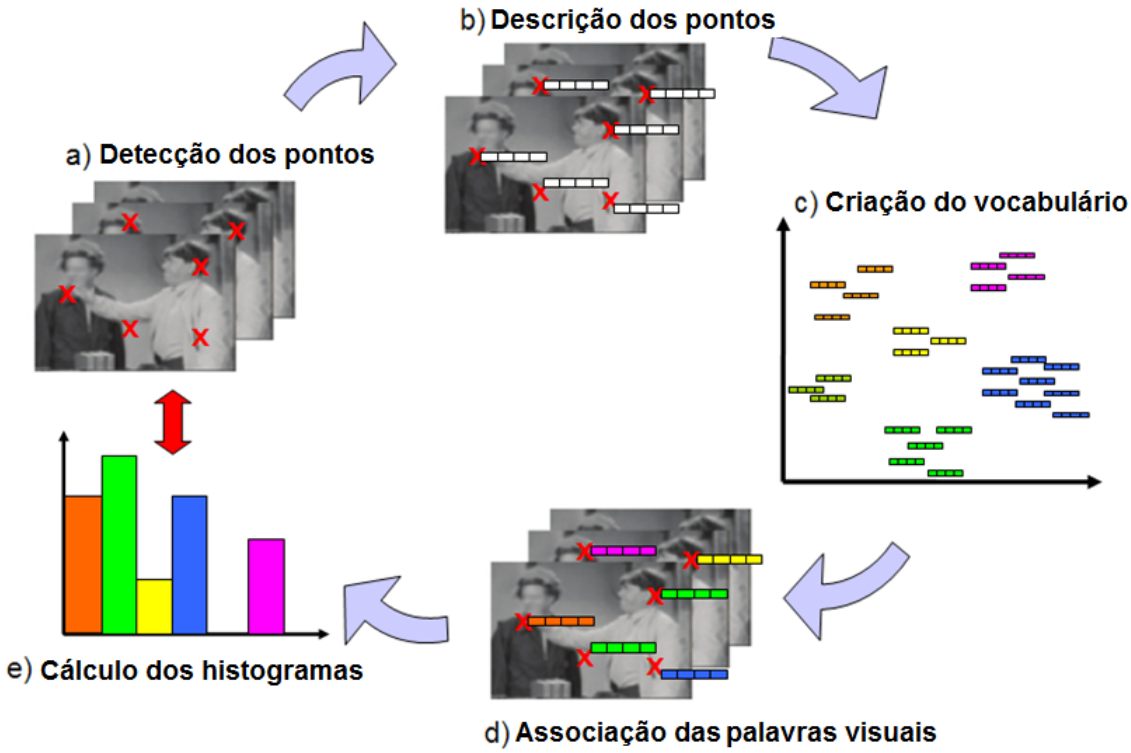

Figura 3.6. Passos para o cálculo dos histogramas de palavras visuais (Lopes et al. [2009]).

Inicialmente são detectados e descritos os pontos de interesse das imagens/segmentos de vídeos, o próximo passo é a criação do vocabulário, onde pontos de interesse representativos (palavras visuais) são escolhidos via agrupamento ou aleatori-

amente como pontos que descrevem padrões mais representativos dentre os detectados. Por fim, é feita a associação das palavras visuais em relação às palavras contidas no vocabulário, essa associação é feita calculando a similaridade entre os pontos detectados e os pontos contidos no vocabulário, então são calculados os histogramas de ocorrências das palavras visuais (Lopes et al. [2009]).

Para a descrição dos segmentos do vídeo foram aplicados os descritores STIP, SIFT e HueSIFT. A seguir é feita uma breve descrição dos descritores considerados neste trabalho.

- O SIFT (*Scale Invariant Feature Transform*) (Lowe [2004]) é um algoritmo para detecção e descrição de pontos de interesse, que objetiva a busca por pontos que apresentam invariância de localização, escala e rotação, além de robustez para transformações geométricas, assim como mudanças de iluminação. Estas características fizeram que o SIFT obtivesse bastante sucesso para o reconhecimento de objetos e o tornou um candidato a aplicações em vídeos. O SIFT é um dos métodos mais utilizados em abordagens de histogramas de palavras visuais. Esse método transforma as informações de uma imagem em um conjunto de coordenadas relativas às características locais, razoavelmente invariantes à escala, mudanças de iluminação, rotação, perspectiva e ruído. O SIFT faz um pré-processamento na imagem, que é o responsável pela identificação dos pontos de interesse. É utilizada a verificação de pontos de máximo e mínimo para a função de diferença de gaussianas-DoG (*Difference-of-Gaussian*). Para cada ponto de interesse identificado é gerado um vetor de características que descreve a região onde o ponto está localizado (Lowe [2004]).

- O HueSIFT (van de Sande et al. [2010]) foi desenvolvido como sendo uma extensão do SIFT que, além da informação do descritor SIFT adiciona informação de cor no espaço HSV. É feita uma concatenação do histograma do canal H. O descritor de Hue-SIFT é invariante à escala e deslocamentos. No entanto, apenas a componente SIFT deste descritor é invariante à iluminação, mudanças de cor ou deslocamento. Resultados experimentais, mostram que para determinadas tarefas, por exemplo classificação de gêneros de vídeos (Lopes et al. [2009]), o HueSIFT obtém melhores resultados que o SIFT, comprovando assim a relevância da informação de cor para essas tarefas.

- O STIP (*Space-Time Interest Point*) (Laptev [2005]) é um algoritmo de detecção de pontos de interesse 3D. O método é uma extensão do detector de Harris (Harris & Stephens [1988]). Ele detecta estruturas locais no espaço e tempo

onde os valores da imagem possuem grande variação local em espaço e tempo. A idéia foi estender o conceito de pontos de interesse do domínio espacial, analisando os valores da imagem em um volume espaço-temporal local para obter grandes variações ao longo das direções espacial e temporal. Pontos com essa característica irão corresponder a pontos de interesse espaciais com localização distinta no tempo correspondendo a uma vizinhança espaço-temporal local sem movimentação constante (Laptev [2005]).

Para a utilização dos descritores SIFT e HueSIFT foi necessário adotar uma estratégia de representação 2D das subtomadas do vídeo. Como todos os quadros de uma mesma subtomada possuem grande similaridade visual, foi extraído o quadro central para representar a subtomada e viabilizar a aplicação dos descritores 2D, assim como feito em Le & Satoh [2007].

Assim, as subtomadas são descritas pelo STIP e os quadros-chave extraídos de cada subtomada são descritos pelo SIFT e pelo HueSIFT. Após a descrição das unidades básicas do vídeo, são computados os BoVFs referentes a todos os segmentos de cada descritor.

São formados três fluxos de processamento como visto na Figura 3.2. O próximo passo do método é o agrupamento dos vetores de palavras visuais referentes a cada descritor. Para calcular a similaridade entre BoVFs foi utilizada a distância Euclidiana.

3.3 Agrupamento

Para a remoção de redundância entre os segmentos do vídeo é aplicado um algoritmo de agrupamento. A ideia é que após o processo de agrupamento, os histogramas de palavras visuais que representam segmentos semelhantes façam parte dos mesmos grupos, e que de cada grupo seja escolhido somente um único segmento como representante.

Após a formação dos BoVFs, é aplicado o algoritmo de agrupamento *k-means* (MacQueen [1967]) em cada um dos fluxos de processamento. Esse método necessita que seja definido o número de grupos k. O número k é definido baseado no número de tomadas do vídeo, que fornece uma estimativa das mudanças de conteúdo visual e o número de cenas presentes no vídeos, uma vez que em *rushes* cada tomada corresponde à gravação de uma cena.

O *k-means* foi adotado por ser uma abordagem simples, bastante difundida e por fornecer bons resultados no processo de classificação não-supervisionada de dados (Richard O. Duda [2001]).

Para a escolha do segmento mais representativo por grupo, é calculada a proximidade dos histogramas pertencentes a um grupo em relação ao seu centroide, o histograma de maior proximidade é escolhido como o representante do grupo. Para o cálculo da proximidade foi utilizada a distância euclidiana. Na Figura 3.7 é visto o processo de agrupamento e escolha do segmento mais representativo por grupo.

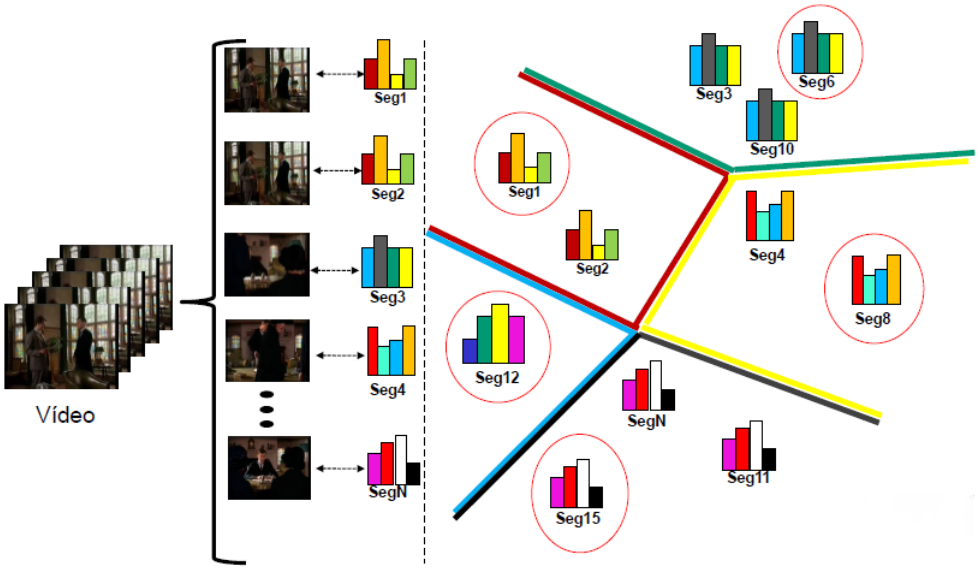

Figura 3.7. Processo de agrupamento.

Uma vez que o número de k não é um valor ótimo, e que podem ocorrer erros no processo de agrupamento, foi incorporada ao método mais uma etapa de eliminação de redundância. Após o processo de agrupamento e da escolha dos segmentos mais representativos, é feita uma comparação entre todos os segmentos selecionados e caso a diferença entre dois segmentos seja menor que um limiar definido previamente ($\epsilon_{similar}$), o segmento de menor valor de movimentação é eliminado.

Para a comparação dos segmentos é utilizada a média dos valores de diferença, $Media_{\chi^2}$, de χ^2.

$$Media_{\chi^2} = \frac{1}{16}\sum_{i=0}^{15}\chi^2(H_s^k[i], G_s^k[i]). \tag{3.4}$$

3.4 Seleção dos segmentos mais informativos

Após as fases anteriores, para gerar o sumário precisamos fazer uma seleção dos segmentos mais significativos. Parte-se da suposição que quanto maior o nível de atividade no segmento, maior é a quantidade de informação fornecida. Nesse trabalho, assim

como nos trabalhos de Pan et al. [2007], Laganière et al. [2008] e Putpuek et al. [2008], utiliza-se medidas de movimentação para o cálculo do nível de atividade. Assim, a média das magnitudes dos vetores de movimentação entre os quadros de um segmento é utilizada como medida de informatividade.

Porém, visto que o cálculo dos vetores de movimentação é uma tarefa de alto custo computacionalmente, é utilizada uma estratégia que se beneficia das características da segmentação em subtomadas para diminuir o tempo de processamento. Uma vez que os quadros de uma mesma subtomada apresentam grande similaridade visual, os vetores de movimentação não são calculados para todos os quadros de um segmento, mas sim entre quadros com intervalo de um segundo.

A suposição é que não haverá grandes modificações entre os quadros de uma subtomada, o que não influenciaria no cálculo dos vetores de movimentação, ao mesmo tempo, exige um tempo de processamento bastante inferior ao necessário para calcular os vetores de movimentação entre todos os quadros da subtomada.

Para garantir que todos os resumos estejam de acordo com um tamanho máximo predefinido, o problema foi modelado como o amplamente conhecido problema da mochila binária (Cormen et al. [2001]).

O problema da mochila (*Knapsack problem*) constitui uma classe de problemas das mais estudadas em otimização combinatória e em subproblemas de outros problemas práticos. O nome refere-se ao modelo de uma situação em que é necessário carregar uma mochila de capacidade limitada com um conjunto de objetos de pesos e valores diferentes. O objetivo é ocupar a mochila com o maior valor possível, não ultrapassando o seu peso máximo. Definir o subconjunto de objetos cujo peso não ultrapasse o limite da mochila e ao mesmo tempo maximizando o seu valor total, corresponde a resolver o problema da mochila (Cormen et al. [2001]).

Dados um conjunto de n objetos e uma mochila com:

- c_j = benefício do objeto j.
- w_j = peso do objeto j.
- b = capacidade da mochila.

Determinar quais objetos devem ser colocados na mochila para maximizar o benefício total de tal forma que o peso da mochila não ultrapasse sua capacidade. Definindo formalmente nosso objetivo é:

$$\begin{cases} maximizar\ z = \sum_{j=1}^{n} c_j s_j \\ Sujeita\,a\ \sum_{j=1}^{n} w_j s_j \leq b \\ s_j \in \{0,1\} \end{cases}$$

Neste trabalho os resumos serão formados pela união dos segmentos que possuam o maior valor de movimentação e que essa união ainda esteja de acordo com o tamanho máximo preestabelecido para o resumo. Assim, foi definido que o número de quadros do resumo seria correspondente ao peso da mochila e o valor de movimentação correspondente ao benefício. Com a resolução do problema da mochila obtêm-se os segmentos que unidos obedecem a um limite de tempo predefinido, ao mesmo tempo que possuem o valor máximo de movimentação.

Para a resolução do problema da mochila foi utilizado o algoritmo baseado em programação dinâmica. Esse método é aprimorado para utilizar recursividade ao invés de executar uma função diversas vezes. Nas primeiras chamadas da função os resultados são armazenados e nas demais chamadas o retorno será o resultado que foi armazenado e não uma requisição à ser resolvida, assim elimina-se a repetição de cálculos, tornando o método mais eficiente. Nesse método, uma solução ótima é definida por uma sequência de decisões ótimas locais. Na programação dinâmica é possível avaliar todas as soluções, garantindo um resultado exato (Cormen et al. [2001]).

3.5 Integração dos resumos

Como uma das contribuições desse trabalho é o uso de vários descritores para a caracterização dos vídeos, o último passo para a geração dos resumos do vídeo consiste em integrar os segmentos que fazem parte dos resumos associados a cada descritor. A ideia de integrar os resumos é inspirada na técnica de aprendizado multivisão (*Multi-view learning*) ou aprendizado multimodal (**?**).

Aprendizado multivisão é um tipo de aprendizado de máquina que explicitamente explora vários conjuntos disjuntos de características, cada uma suficientemente capaz para se aprender um conceito alvo. A idéia é que as características se complementem e gerem melhores resultados dos que os obtidos com uma única descrição.

No aprendizado multivisão um dado é descrito em mais de uma maneira e essas descrições recebem o nome de visões. Idealmente, as visões devem ser disjuntas e não correlatas, mas é visto na literatura que na prática isso é difícil de ser alcançado. Para cada visão é treinado um classificador, no final, os resultados gerados pelos classifica-

dores são combinados objetivando a diminuição no erro de classificação (Muslea et al. [2002]).

Assim, foi criada uma configuração que utiliza aprendizado não-supervisionado para a classificação dos segmentos de vídeo. Nessa configuração as características geradas por cada descritor são as visões e os classificadores são os algoritmos de agrupamento. Por fim, os resumos gerados para cada descritor são combinados.

Após a união dos resumos, caso haja sobreposição entre dois segmentos, é formado um novo segmento com o menor valor inicial e o maior valor final das subtomadas envolvidas na sobreposição, e é calculado um novo valor de movimentação para esse segmento.

Em seguida, são repetidos os mesmos passos para a formação dos resumos associados a cada descritor. É feita a comparação dos segmentos para eliminação de redundância, depois são selecionados os segmentos mais representativos de acordo com o algoritmo de resolução do problema da mochila. Por último os segmentos selecionados são organizados em ordem temporal, formando o resumo do vídeo.

Capítulo 4

Resultados Experimentais

Neste capítulo, são apresentados os experimentos desenvolvidos para avaliar os sumários de vídeos gerados pelo presente trabalho. É feita uma breve descrição da base de dados utilizada, e dos trabalhos escolhidos para a comparação do método proposto. Em seguida, é descrita a metodologia de avaliação feita por usuários. Por fim, os resultados quantitativos e qualitativos obtidos são apresentados e discutidos.

Para a realização dos experimentos alguns parâmetros tiveram que ser estimados experimentalmente, os parâmetros utilizados e seus valores são:

- $\epsilon_{Single} = 4500$.
- $\epsilon_{SubShot} = 3500$.
- $\epsilon_{Cut} = 1000$.
- $\epsilon_{ColorBar} = 300$.
- $\epsilon_{Motion} = 1000000$.
- $\epsilon_{Similar} = 500$.
- Tamanho do vocabulário visual = 1000

No restante do texto serão utilizadas siglas para identificar os resumos de cada sistema. A saber: R-SHS (resumos formados pela união dos resumos do SIFT, Hue-SIFT e STIP), R-SIFT (resumos gerados a partir do descritor Hue-SIFT), R-Hue (resumos gerados a partir do descritor Hue-SIFT), R-STIP(resumos gerados a partir do descritor STIP), CityU (resumos gerados pelo trabalho de Wang et al. [2007]), Lip6 (resumos gerados pelo trabalho de Detyniecki & Marsala [2007]) e Nii (resumos gerados pelo trabalho de Le & Satoh [2007]).

4.1 Base de dados

A base de vídeos utilizada no presente trabalho é a mesma utilizada na tarefa de sumarização do TRECVID 2007 (Over et al. [2007]). Essa conformidade visa viabilizar a comparação com outros trabalhos submetidos à competição. A base de vídeos consiste de material não editado filmado para cinco séries dramáticas da BBC. Os vídeos foram cedidos pela BBC *Archive*, e mostram pessoas em várias situações do cotidiano, tanto em ambientes fechados quanto em ambientes abertos. Existem sons de diálogos e também sons naturais do diretor, da equipe de produção e do ambiente de filmagem. Existe também muita redundância, pois as cenas foram gravadas e regravadas em vários ângulos. A equipe de produção aparece em alguns momentos, assim como barras de cor e claquetes nas fronteiras de cenas e tomadas (Over et al. [2007]).

Aproximadamente 50 vídeos foram providos aos grupos participantes da tarefa de sumarização do TRECVID como dados de desenvolvimento, enquanto 42 foram mantidos para uso em testes nos sistemas desenvolvidos. Cada conjunto de vídeos é uma amostra aleatória balanceada com respeito ao número de vídeos de cada série da BBC. Um conjunto de metadados representando anotações (*ground truth*) de exemplo também foi disponibilizado. As anotações consistem em uma lista que representa importantes segmentos dos vídeos contendo descrições de objetos e ocorrência de eventos. Nas anotações há sempre uma preocupação em especificar ângulo de câmera, distância ou outras informações que tornem o segmento único.

Os vídeos têm uma duração mínima de 3,3 minutos e uma duração máxima de pouco menos de 37 minutos, com média de duração de 25 minutos. A Figura 4.1 apresenta a distribuição das durações dos 42 vídeos, a Figura 4.2 apresenta a duração alvo dos resumos que corresponde a 4% da duração do vídeo original, como definida na competição do TRECVID 2007 (Over et al. [2007]).

4.2 Abordagens comparadas

Para comparação do presente trabalho com a literatura, foram selecionados os três trabalhos que obtiveram as melhores taxas de inclusão de itens do *ground truth* na tarefa de sumarização do TRECVID 2007. A avaliação do TRECVID 2007 não mostrou diferença estatística entre os três trabalhos em relação à taxa de inclusão. Os trabalhos são: Wang et al. [2007], Detyniecki & Marsala [2007] e Le & Satoh [2007].

Wang et al. [2007] (CityU) desenvolveram uma abordagem de sumarização baseada em uma complexa e detalhada análise do vídeo original, que inclui detecção de tomadas, detecção de objetos, estimativa de movimento de câmera, correspondência e

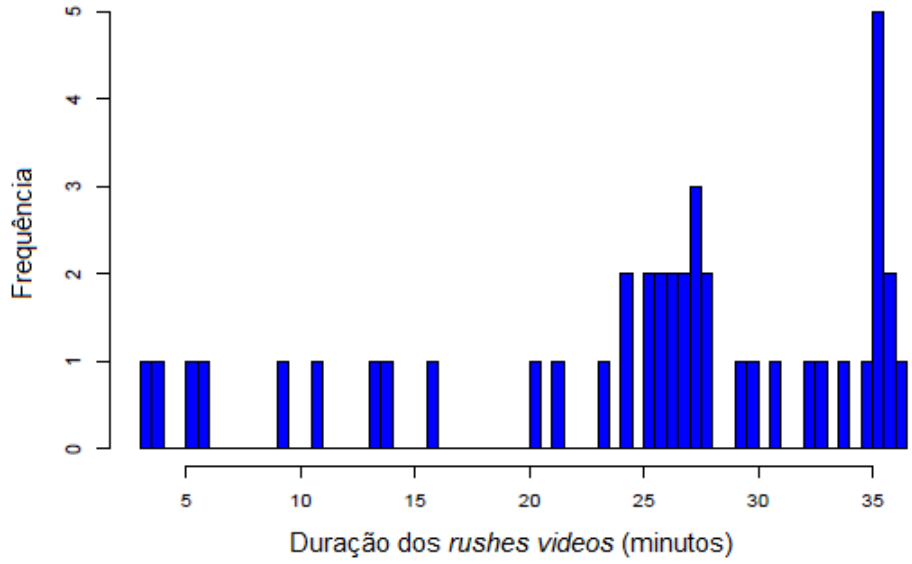

Figura 4.1. Distribuição da duração dos vídeos. (Over et al. [2007])

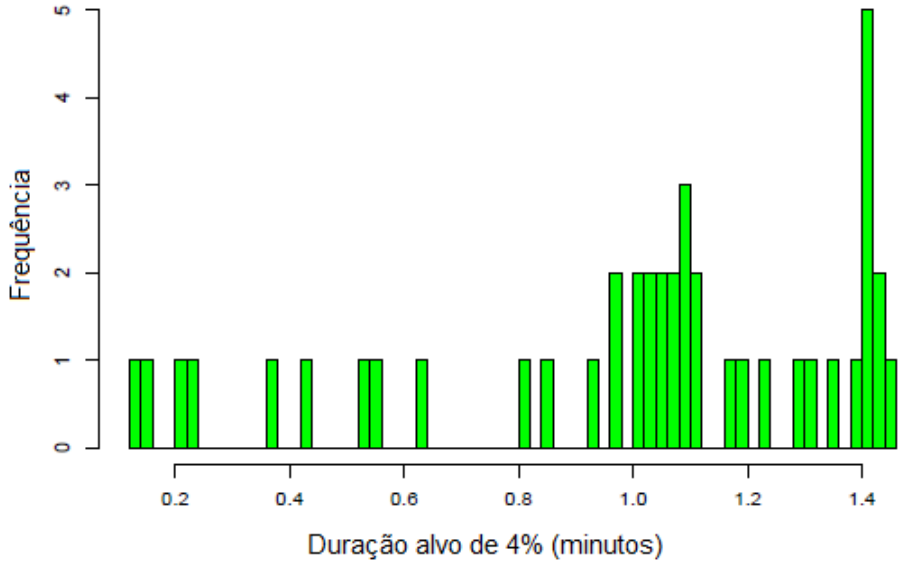

Figura 4.2. Distribuição da duração alvo dos resumos. (Over et al. [2007])

rastreamento de regiões de interesse, classificação de áudio e reconhecimento de fala. Uma medida de representatividade foi utilizada para modelar a presença de objetos e quatro eventos audiovisuais (movimentação de objetos, movimento de câmera, mudança de cena e conteúdo de fala) em segmentos de vídeos. Os segmentos com maior representatividade foram selecionados para formarem o resumo. A detecção de *junk*

shots é realizada comparando as características de cor dos segmentos com as características de cor de modelos padrões de *junk shots* (por exemplo, barras de cor e quadros de cor única), os autores não especificam se essa etapa de detecção inclui a detecção de claquetes.

Já Detyniecki & Marsala [2007] (Lip6) desenvolveram uma abordagem para sumarização de vídeos baseada em detecção de tomadas e eliminação de conteúdo redundante através de uma técnica que eles chamaram de *stacking*. Nessa técnica as tomadas são comparadas e dentre as consideradas semelhantes somente a de maior duração é selecionada. O próximo passo na construção do resumo é a aplicação de uma técnica de aceleração adaptativa dos quadros do vídeo. É definida uma medida de importância baseada nas semelhança visual entre os quadros das tomadas selecionadas, então as tomadas mais importantes são exibidas a uma taxa normal de quadros por segundo, enquanto as tomadas consideradas não informativas são expostas com uma taxa maior de quadro por segundos. Um dos problemas desse trabalho é o fato de não ter sido realizada a detecção de *junk shots*.

Le & Satoh [2007] (Nii) propuseram uma abordagem baseada em características locais de cor e agrupamento para a sumarização de rushes vídeos. Inicialmente o vídeo é dividido em segmentos que apresentam grande similaridade visual, em seguida o quadro central de cada segmento é extraído como quadro representativo. Os quadros representativos são agrupados e de cada grupo é selecionado o segmento que possua maior duração. Por último, é feito um cálculo baseado na duração do sumário para a seleção de amostras dos segmentos. Esse trabalho também possui o problema de não realizar a detecção de *junk shots*

Esses trabalhos foram comparados com os quatro tipos de resumos gerados na metodologia proposta.

4.3 Metodologia de avaliação

Os métodos de sumarização automática de vídeos devem ser avaliados para verificar a relevância dos segmentos de vídeo selecionados bem como, para o caso específico de *rushes videos*, avaliar se a detecção de *junk shots* foi efetiva. Apesar do problema de sumarização estar sendo intensamente investigado, não há um método ideal para avaliar a qualidade dos resumos produzidos (Money & Agius [2008]). Alguns esforços foram feitos para criar um método padronizado, como o proposto na tarefa de sumarização de vídeos do TRECVID 2007 (Over et al. [2007]).

Neste trabalho, os resumos são avaliados de maneira subjetiva, através de usuá-

rios. A metodologia de avaliação é baseada na metodologia empregada na tarefa de sumarização do TRECVID 2007.

O TRECVID 2007 definiu um procedimento de avaliação em que avaliadores humanos assistem aos resumos para avaliar a quantidade de conteúdo relevante que eles contêm. Foi fornecido aos participantes uma base de vídeos anotados em que o *ground truth* é uma lista de tópicos que representa importantes segmentos do vídeo que deveriam estar contido em um bom sumário. Nas anotações há a preocupação em especificar o ângulo da câmera, distância ou outras informações que tornem o segmento único.

Cada resumo produzido foi julgado por três juízes humanos. Cada avaliador recebeu o resumo e uma lista correspondente a até 12 tópicos do *ground truth* selecionados aleatoriamente. O avaliador assiste ao resumo a 25 quadros por segundo usando apenas controles de tocar e pausar. Em seguida, ele tenta identificar os tópicos do *ground truth* que apareceram no resumo. O percentual de tópicos encontrados por cada avaliador determina a fração de segmentos importantes do vídeo original que foram incluídos no resumo.

O processo de tentar encontrar a lista de tópicos no resumo é cronometrado para computar uma medida de esforço no julgamento. O processo de avaliação também coleta medidas de usabilidade/satisfação. Para avaliar a quantidade de segmentos redundantes e *junk shots* presentes no sumário foi definida uma escala variando de 1 a 5, que vai de concordo fortemente a discordo fortemente.

Uma descrição completa da metodologia de avaliação é encontra em Over et al. [2007]. No TRECVID 2007 as medidas usadas para cada sumário foram:

- Porcentagem de tópicos encontrados pelo avaliador;
- Facilidade em encontrar o conteúdo desejado;
- Quantidade de segmentos redundantes e *junk shots* presentes no resumo;
- Quantidade de tempo gasto para a avaliação do sumário;
- Tamanho do sumário, relativo aos 4% do tamanho do vídeo original;
- Tempo que o sistema levou para criar o sumário;

A pontuação total de um resumo é a média das pontuações atribuídas pelos três avaliadores. A Figura 4.3 mostra o processo de avaliação de um resumo. Os resultados da avaliação manual foram analisados estatisticamente e chegou-se a conclusão

que houve uma boa concordância entre os julgamentos dos avaliadores com base na comparação dos tópicos detectados por dois avaliadores em um resumo.

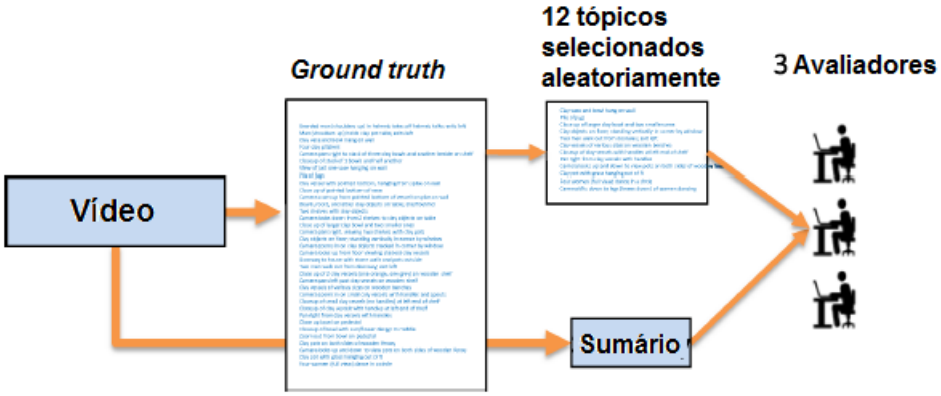

Figura 4.3. Processo de avaliação do TRECVID 2007. (Dumont & Mérialdo [2010])

Como a avaliação por usuários é uma tarefa que requer muito esforço e tempo, pois deve-se conseguir uma equipe para a avaliação, foi decido que somente algumas das métricas utilizadas no TRECVID 2007 seriam utilizadas para a avaliação dos resumos produzidos nesse trabalho. O foco foi dado para a porcentagem de tópicos do *ground truth* encontrados no resumo, pois entende-se que essa **métrica objetiva** é a medida mais importante para avaliar a qualidade de um resumo. Outras **métricas subjetivas** como quantidade de segmentos redundantes presentes no vídeo, quantidade de *junk shots* e tempo gasto na avaliação, também foram calculadas.

Os resumos foram apresentados agrupados por vídeo a ser sumarizado. Dentro de cada grupo, os resumos foram apresentados em ordem aleatória para tentar diminuir qualquer viés de aprendizado de contexto. Antes de julgar um grupo de resumos, foi recomendado a cada avaliador que estudasse a lista de tópicos para o presente resumo.

Para realizar a avaliação dos resumos foram selecionados 42 avaliadores, todos provenientes da área de ciência da computação. Os usuários foram divididos em grupos de três, onde cada grupo ficou responsável pela avaliação de três vídeos. Cada componente de um grupo avaliou os resumos referentes a três vídeos. Essa abordagem está de acordo com a avaliação do TRECVID 2007, onde todos os resumos foram avaliados por três julgadores e os resultados foram dados em função da média das três avaliações.

Antes de iniciar a avaliação, as seguintes instruções foram fornecidas aos usuários:

- **Passo 1:** Fazer uma leitura de todos os conceitos à direita do *player* de vídeo, se familiarizando com o contexto do vídeo original (esses conceitos são itens do

ground truth, note que o avaliador não assiste ao vídeo original);

- **Passo 2:** Clicar no botão iniciar para habilitar a lista de conceitos e as escalas;
- **Passo 3:** Iniciar o vídeo (é permitido que sejam feitas pausas, retrocessos ou que o vídeo seja assistido mais de uma vez);
- **Passo 4:** Marcar os conceitos que foram identificados no vídeo;
- **Passo 5:** Marcar nas escalas abaixo do *player* a sua opinião sobre a quantidade de trechos redundantes e sobre a quantidade de *junk shots* presente no vídeo. Essas escalas têm 5 níveis que variam de 1 (concordo fortemente) a 5 (discordo fortemente);
- **Passo 6:** Clicar no botão encerrar, finalizando a avaliação para o resumo atual;

A Figura 4.4 mostra o sistema utilizado para a avaliação dos resumos. Para facilitar o processo de avaliação, o sistema foi desenvolvido numa plataforma *web*, evitando assim que os avaliadores necessitassem ir a um local específico para realizar a avaliação. Para cada usuário são disponibilizados 21 resumos referentes a três vídeos e sete técnicas (R-SHS, R-SIFT, R-Hue, R-STIP, CityU, Lip6 e Nii). Esses resumos são apresentados ao avaliador agrupados por vídeo e em ordem aleatória em relação ao método gerador.

4.4 Análise dos resultados

Determinar a razão de um sistema obter bons resultados para um vídeo específico e resultados ruins para um outro vídeo é uma tarefa complexa. Dentre os fatores que podem influenciar nessa conclusão estão o conteúdo do vídeo, o avaliador e o *ground truth*. Após analisar esses três fatores e os valores de inclusão para cada vídeo, chegou-se a conclusão que o principal fator de influência é o conteúdo do vídeo envolvido na avaliação.

Vídeos com maior quantidade de cenas, cenários, com a presença dos atores bem definida tendem a obter taxa de inclusão mais altas. Por outro lado, vídeos com número limitado de cenas, pouca variação de cenários e com poucos atores tornam-se mais confusos. De fato, sistemas que utilizem o mecanismo de geração dos resumos por eliminação de redundância possuem grande dificuldade em gerar sumários concisos para esse tipo de vídeo.

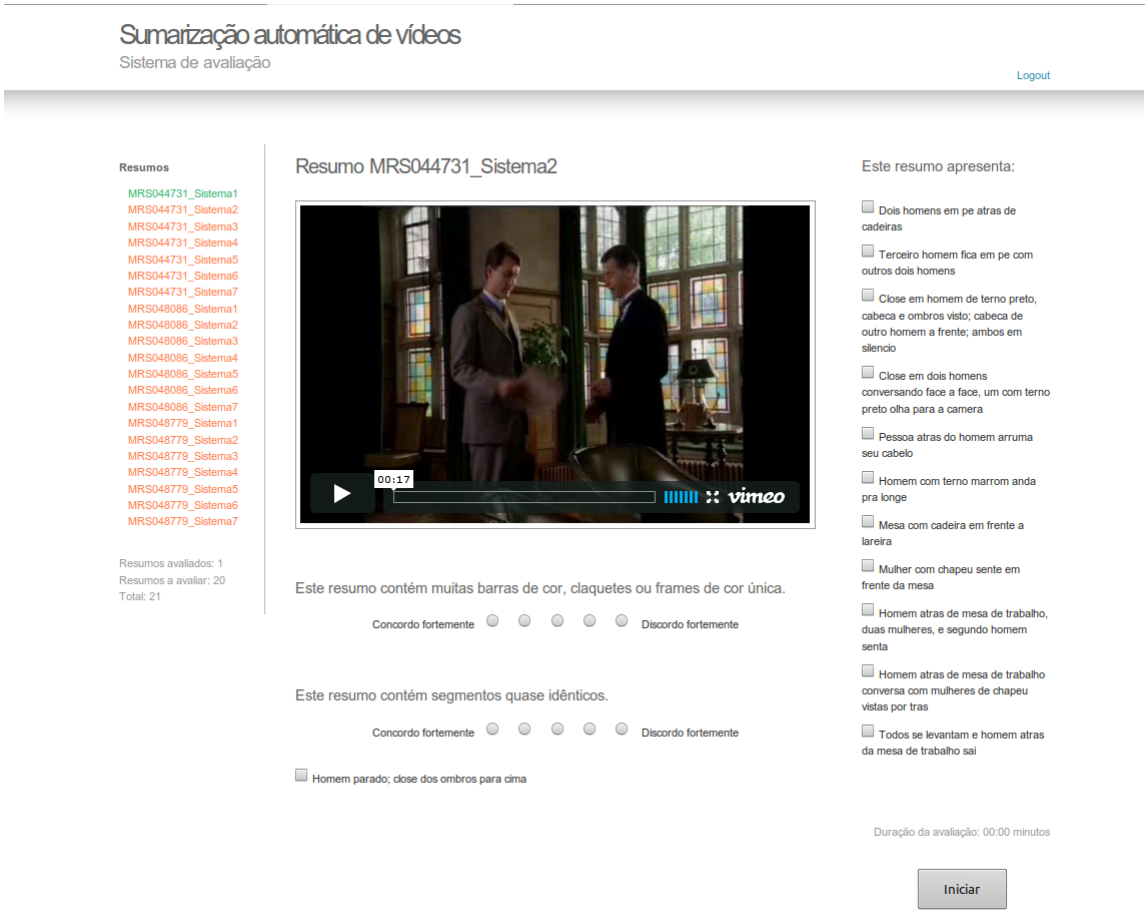

Figura 4.4. Sistema de avaliação dos resumos.

Nesta seção é apresentada uma análise exploratória dos resultados produzidos pela avaliação dos sete sistemas (R-SHS, R-SIFT, R-Hue, R-STIP, CityU, Lip6 e Nii). Os resultados são apresentados em forma de gráficos de caixas, e um teste de hipótese é feito para comparar os valores obtidos pelos sistemas em relação à cada medida avaliada.

Como teste de hipótese foi utilizado o *t-teste* (Jain [1991]), com confiança de 1 - α (neste trabalho 1 - $\alpha = 0{,}95$). As hipóteses são enunciadas da seguinte forma:

$$H_0 : \text{Média 1} = \text{Média 2 vs. } H_1 : \text{Média 1} > \text{Média 2.} \tag{4.1}$$

a Equação 4.1 faz a comparação entre os resultados alcançados por dois sistemas. Média 1 e Média 2 representam as médias dos valores alcançados para a medida avaliada. A hipótese nula, H_0, é tida como verdadeira até que provas estatísticas indiquem o contrário. H_0 afirmar que os resultados para a medida avaliada são estatisticamente equivalentes. Por outro lado, a hipótese alternativa, H_1, afirma que a média observada

para o sistema 1 é maior que a média do sistema 2. Tradicionalmente, H_0 é rejeitada caso o p-valor[1] seja menor ou igual ao nível de significância α (neste trabalho definido como 0,05) (Jain [1991]).

Nas próximas subseções os testes de hipótese comparam os resumos gerados pelo presente trabalho entre si e com os três melhores colocados da tarefa de sumarização do TRECVID 2007.

4.4.1 Inclusão dos itens do *ground truth*

O valor de inclusão dos itens do *ground truth* pode variar de 0 a 100, com uma granularidade máxima de 8,33 (100/12, onde 12 corresponde à quantidade de itens do *ground truth* avaliados). A Tabela 4.1 mostra a média dos valores alcançados para todos os sumários de cada sistema. Os valores variaram de 49,06 (Lip6) a 58,63 (Nii). Esses valores mostram a grande subjetividade do problema de sumarização, pois apesar do sistema Nii fazer uma análise em baixo nível do vídeo obtém os melhores resultados de média, inicialmente esperava-se que o sistema CityU obtivesse os melhores resultados por fazer uma análise bem detalhado do vídeo, porém isso não foi observado. Os piores resultados são alcançados pelo sistema Lip6 que escolhe os segmentos que irão compor o resumo por similaridade visual e aplica uma aceleração adaptativa, que como exposto no Capítulo 2, pode tornar algumas cenas difíceis de compreender.

Em relação aos sistemas da abordagem proposta, foram alcançados bons resultados. O sistema R-SHS se beneficiou do fato de ser formado pelos segmentos mais informativos da união dos outros três resumos (R-SIFT, R-Hue e R-STIP) e obteve a melhor média, enquanto que os outros sistemas obtiveram valores menores e muito próximos.

Tabela 4.1. Média das taxas de inclusão

	Taxa de inclusão
R-SHS	57,11
R-SIFT	52,76
R-HueSIFT	54,44
R-STIP	53,83
CityU	53,97
Lip6	49,07
Nii	58,63

[1] p-valor é definido como a probabilidade de se obter uma estatística de teste igual ou mais extrema quanto aquela observada em uma amostra, assumindo verdadeira a hipótese nula.

A Figura 4.5 apresenta o gráfico de caixas referente aos valores alcançados por cada sistema para a inclusão dos itens do *ground truth*. As caixas no gráfico representam os intervalos existentes entre o primeiro e o terceiro quartis das amostras, isto é, o local onde se encaixam 50% das medidas obtidas. A linha na caixa indica o valor de mediana dos dados. Se a linha mediana dentro da caixa não é equidistante dos extremos, diz-se então que os dados são assimétricos. Já, os extremos do gráfico indicam os valores mínimo e máximo, a menos que valores *outliers* estejam presentes, nesse caso o gráfico se estende ao máximo de 1,5 vezes da distância inter-quartil. Os pontos fora do gráfico são os *outliers* ou suspeitos de serem *outliers*, por último, a linha tracejada em vermelho representa o intervalo de confiança.

Como há sobreposição do intervalo de confiança de todos os sistemas, então a Figura 4.5 não fornece muita informação a respeito das diferenças estatísticas entre as amostras dos sistemas avaliados. Então, para uma análise mais detalhada sobre as diferenças apresentadas entre as medidas, foi realizado um teste de hipótese.

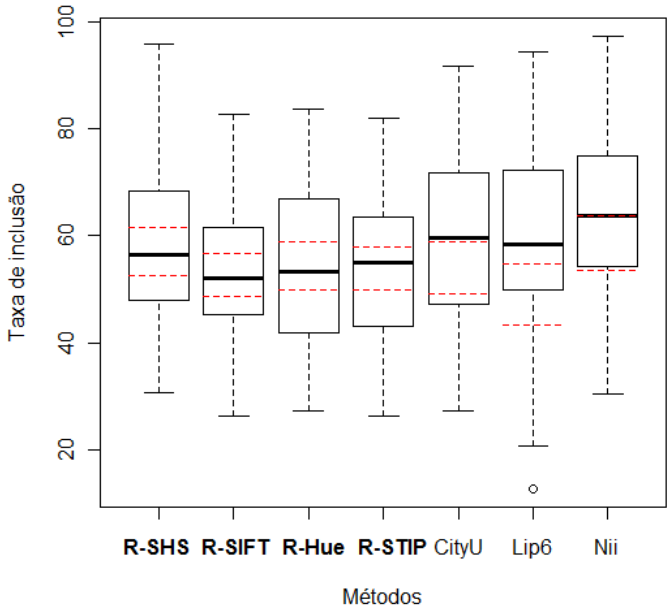

Figura 4.5. Porcentagem de inclusão dos itens do *ground truth*

A Tabela 4.2 mostra os *p*-valores gerados pelo *t-test* quando comparando os métodos da abordagem proposta entre si e com os trabalhos da literatura. Analisando a tabela percebe-se que somente dois valores (R-SHS - R-SIFT e R-SHS - R-STIP) são menores que o nível de significância $\alpha = 0{,}05$, e para esses dois casos a hipótese nula é rejeitada. Assim pode-se concluir, com 95% de confiança, que o sistema R-SHS é

estatisticamente equivalente aos sistemas R-Hue, CityU, Lip6 e Nii, e que é superior estatisticamente quando comparado ao R-SIFT e R-STIP em relação a taxa de inclusão dos itens do *ground truth*. Os outros sistemas mostraram-se equivalente entre si e em relação aos trabalhos da literatura.

Tabela 4.2. *T-test* para a medida de inclusão dos itens do *ground truth*

	p-valor
R-SHS - R-SIFT	0,02
R-SHS - R-Hue	0,08
R-SHS - R-STIP	0,04
R-SHS - CityU	0,5
R-SHS - Lip6	0,5
R-SHS - Nii	0,8
R-SIFT - R-Hue	0,7
R-SIFT - R-STIP	0,6
R-SIFT - CityU	0,9
R-SIFT - Lip6	0,9
R-SIFT - Nii	0,4
R-Hue - R-STIP	0,9
R-Hue - CityU	0,9
R-Hue - Lip6	0,9
R-Hue - Nii	0,9
R-STIP - CityU	0,9
R-STIP - Lip6	0,9
R-STIP - Nii	0,9

4.4.2 Presença de *junk shots*

Para a avaliação da presença de *junk shots* nos resumos, foi definida uma escala que varia de 1 (concordo fortemente) a 5 (discordo fortemente) com a presença de *junk shots*. A Tabela 4.3 mostra as médias os resultados alcançados por cada sistema. Ela mostra que os resumos produzidos no presente trabalho, bem como o resumo gerado pelo CityU, tiveram melhores resultados que os resumos gerados pelo Lip6 e Nii. Esse é um resultado já esperado, devido ao fato dos sistemas Lip6 e Nii não realizarem a detecção de *junk shots*.

A Figura 4.6 mostra o gráfico de caixas para as medidas obtidas em relação à presença de *junk shots*. O fato de não existir sobreposição dos intervalos de confiança dos sistemas Lip6 e Nii com os demais intervalos de confiança, garante que há diferença significativa dos dois sistemas em relação aos demais.

Tabela 4.3. Médias das medidas de presença de *junk shots*

	Junk shots
R-SHS	4,28
R-SIFT	4,23
R-Hue	4,29
R-STIP	4,31
CityU	4,50
Lip6	2,60
Nii	2,27

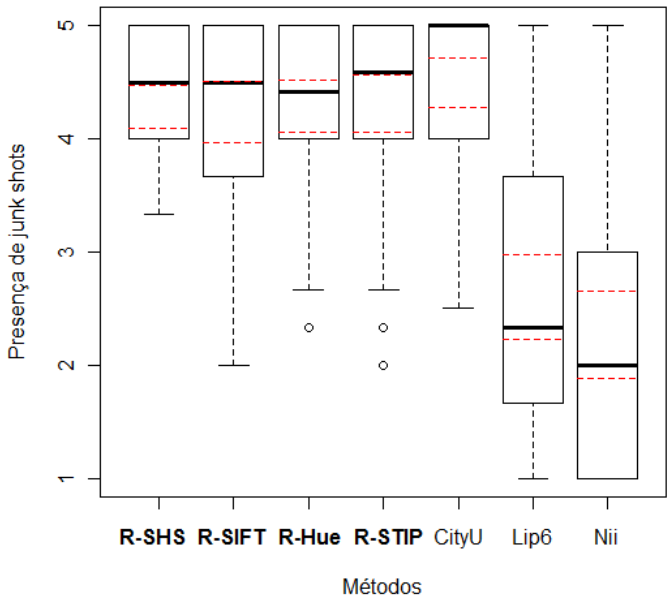

Figura 4.6. Presença de *junk shots*

A análise da Tabela 4.4 mostra que todos os p-valores gerados para as comparações envolvendo os sistemas Lip6 e Nii são menores que α, o que confirma, com 95% de confiança, o que foi visto na Tabela 4.4 e Figura 4.6. Os sistemas Lip6 e Nii são estatisticamente inferiores quando comparados com os outro sistemas em relação à presença de *junk shots*. Já os sistemas da abordagem proposta mostraram-se estatisticamente equivalentes entre si e ao sistema CityU.

Tabela 4.4. *T-test* para a medida de presença de *junk shots*

	p-valor
R-SHS - R-SIFT	0,1
R-SHS - R-Hue	0,2
R-SHS - R-STIP	0,3
R-SHS - CityU	0,7
R-SHS - Lip6	1e-8
R-SHS - Nii	4e-10
R-SIFT - R-Hue	0,6
R-SIFT - R-STIP	0,6
R-SIFT - CityU	0,9
R-SIFT - Lip6	8e-5
R-SIFT - Nii	7e-7
R-Hue - R-STIP	0,5
R-Hue - CityU	0,9
R-Hue - Lip6	4e-6
R-Hue - Nii	2e-9
R-STIP - CityU	0,8
R-STIP - Lip6	2e-6
R-STIP - Nii	2e-7

4.4.3 Presença de segmentos redundantes

Para a avaliação da presença de segmentos redundantes, mais uma vez foi definida uma escala variando de 1 (concordo fortemente) a 5 (discordo fortemente). A Tabela 4.5 apresenta as médias alcançadas por cada sistema, enquanto que a Figura 4.7 mostra os gráficos de caixas, onde os resumos gerados pelo CityU apresentaram menor quantidade de segmentos redundantes em relação aos outros sistemas, enquanto que os demais sistemas obtiveram métricas muito similares entre si. Nas avaliações do TRECVID em 2007, essa medida mostrou-se relacionada à taxa de inclusão dos itens do ground truth, pois a presença de segmentos idênticos pode facilitar a visualização de conteúdo do *ground truth* e, consequentemente, aumentar as taxas de inclusão. Essa tendência não foi observada na avaliação da metodologia proposta, pois o sistema que alcançou os melhores valores de média, para a inclusão do conteúdo do *ground truth*, também obteve os melhores resultados para a ausência de segmentos redundantes.

A Tabela 4.6 apresenta os p-valores gerados para a comparação das médias dos valores para a presença de segmentos redundantes. Nenhum dos p-valores é menor que o nível de significância α, porém ao analisar os valores para as comparações envolvendo o sistema CityU, percebe-se que ao realizar o teste de hipótese unilateral a esquerda (hipótese alternativa: média-CityU $>$ média) pode-se afirmar que, com 95% confiança,

Tabela 4.5. Médias das medidas de presença de segmentos redundantes

	Segmentos redundantes
R-SHS	2,72
R-SIFT	2,84
R-Hue	2,83
R-STIP	2,88
CityU	3,70
Lip6	2,80
Nii	2,75

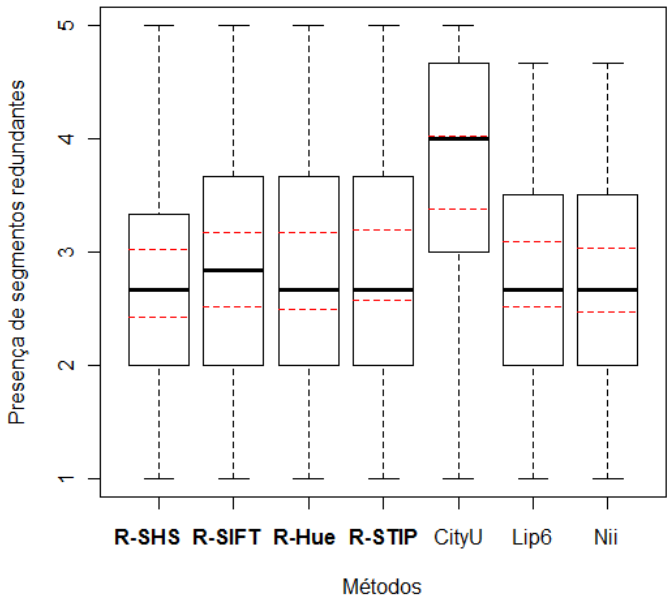

Figura 4.7. Presença de segmentos redundantes

o sistema CityU é estatisticamente superior aos sistemas da metodologia proposta em relação a presença de segmentos redundantes. Isso se explica pelo fato do sistema CityU utilizar uma técnica complexa baseada em busca em grafo e várias características de cor para a remoção de redundância. Enquanto que os outros sistemas utilizam somente similaridade visual baseada em histogramas locais de cor para a remoção de redundância.

Tabela 4.6. *T-test* para a medida de presença de segmentos redundantes

	p-valor
R-SHS - R-SIFT	0,6
R-SHS - R-Hue	0,6
R-SHS - R-STIP	0,7
R-SHS - CityU	0,99
R-SHS - Lip6	0.6
R-SHS - Nii	0,5
R-SIFT - R-Hue	0,5
R-SIFT - R-STIP	0,5
R-SIFT - CityU	0,99
R-SIFT - Lip6	0,43
R-SIFT - Nii	0,3
R-Hue - R-STIP	0,6
R-Hue - CityU	0,99
R-Hue - Lip6	0,4
R-Hue - Nii	0,4
R-STIP - CityU	0,99
R-STIP - Lip6	0,3
R-STIP - Nii	0,3

4.4.4 Tempo de julgamento

Das métricas avaliadas, essa é a mais subjetiva pois é totalmente dependente do usuário. Ela foi escolhida para fazer parte da avaliação por dar uma noção de facilidade do entendimento do resumo e na tarefa de identificação dos itens do *ground truth*. Essa foi a métrica que apresentou a maior quantidade de *outliers* e maior dispersão dos dados. Por ser uma métrica muito subjetiva, não foi possível identificar uma razão clara para esse comportamento em relação ao tempo de avaliação. Esperava-se que os métodos que utilizam aceleração de segmentos necessitassem de uma quantidade maior de de tempo para a avaliação, porém isso não foi observado. Por outro lado, foi identificado que os sistemas que obtiveram as melhores médias de taxa de inclusão também foram os que apresentaram maior tempo de avaliação.

A Tabela 4.7 mostra os valores das médias dos tempos de avaliação para cada sistema. Os valores variaram de 79,61 (R-STIP) a 126,02 (R-SHS)

Analisando a Figura 4.8, os sistema Nii e R-SHS foram os que apresentaram os maiores tempos de avaliação, enquanto que os outros sistemas obtiveram medidas similares.

A Tabela 4.8 mostra que os sistemas R-SHS, R-Hue e R-STIP foram estatisticamente equivalentes quando comparados com os sistemas CityU, Lip6 e Nii. Analisando

Tabela 4.7. Médias das medidas de tempo de julgamento por usuários

	Tempo (em segundos)
R-SHS	126,02
R-SIFT	81,08
R-Hue	92,67
R-STIP	79,61
CityU	109,79
Lip6	82,03
Nii	130,99

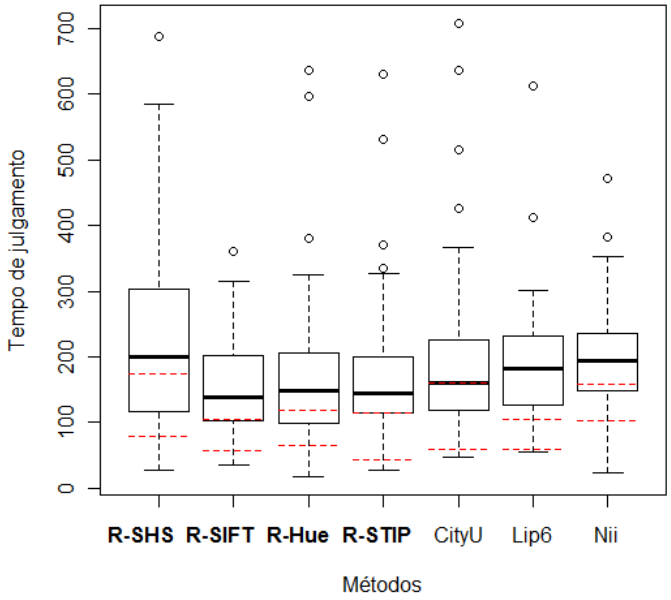

Figura 4.8. Tempo de julgamento por usuários (em segundos)

os valores das comparações para o sistema SIFT, percebe-se que, com 95% de confiança, ele foi estatisticamente inferior ao tempo de avaliação dos sistemas CityU, Lip6 e Nii, superior quando comparado ao sistema R-SHS e equivalente quando comparado aos sistemas R-Hue e R-STIP.

4.5 Considerações

Conforme os resultados experimentais obtidos para as quatro medidas avaliadas, observou-se que a abordagem proposta apresentou resumos com qualidade estatisticamente equivalente às demais abordagens comparadas. Para a taxa de inclusão dos

Tabela 4.8. *T-test* para a medida de tempo

	p-valor
R-SHS - R-SIFT	0,008
R-SHS - R-Hue	0,09
R-SHS - R-STIP	0,07
R-SHS - CityU	0,3
R-SHS - Lip6	0,2
R-SHS - Nii	0,2
R-SIFT - R-Hue	0,8
R-SIFT - R-STIP	0,84
R-SIFT - CityU	0,96
R-SIFT - Lip6	0,98
R-SIFT - Nii	0,99
R-Hue - R-STIP	0,5
R-Hue - CityU	0,78
R-Hue - Lip6	0,77
R-Hue - Nii	0,88
R-STIP - CityU	0,82
R-STIP - Lip6	0,81
R-STIP - Nii	0,90

itens do *ground truth*, o sistema R-SHS mostrou-se, com 95% de confiança, superior aos resumos gerados pelo R-SIFT e R-STIP, comprovando assim que a união dos resumos traz benefícios para a sumarização de vídeos.

O método de detecção de *junk shots* utilizado na metodologia proposta mostrou-se muito robusto e obteve os melhores resultados juntamente com o sistema CityU. Em relação à presença de segmentos redundantes, a abordagem proposta não obteve bons resultados, porém, como observado na avaliação do TRECVID, a presença de segmentos redundantes pode facilitar o entendimento do itens do *ground truth* e consequentemente aumentar a taxa de inclusão. A medida de tempo mostrou as melhores taxas de inclusão foram alcançadas com um maior tempo de avaliação, contudo não foi verificada uma evidência clara que justifique o tempo de avaliação para um determinado sistema.

4.6 Exemplo de sumário

A Figura 4.9 mostra algumas imagens do resumo gerado para o menor vídeo da base. O vídeo original tem uma duração de 5,5 minutos, enquanto o resumo gerado pela união dos descritores possui uma duração de 12 segundos. Os itens do *ground truth* para esse vídeo são:

- Enfermeira fala ao telefone, luz visível através da porta.
- Enfermeira desliga o telefone, luz visível através da porta.
- Visão da enfermeira sem o telefone, luz visível através da porta.
- Enfermeira fala ao telefone, nenhuma luz visível através da porta.
- Enfermeira desliga o telefone, nenhuma luz visível através da porta.
- Visão da enfermeira sem o telefone, nenhuma luz visível através da porta.

Figura 4.9. Imagens retiradas do resumo

Capítulo 5

Conclusões e trabalhos futuros

Este trabalho apresentou uma metodologia para a geração automática de resumos dinâmicos de *rushes videos*. A metodologia proposta integrou vantagens observadas na literatura relacionada, agregando em um único método técnicas de comprovada eficácia na geração de resumos dinâmicos.

Este trabalho concentrou-se na classificação não-supervisionada dos segmentos dos vídeos baseada em informações espaciais e espaço-temporais associadas a uma técnica de histogramas de palavras visuais. Como a representação por histogramas de palavras visuais tem sido utilizada na literatura em vários cenários de detecção e classificação de padrões e provado ser robusta nessas tarefas, optou-se por utilizá-la no problema de sumarização automática de vídeos, com uma implementação baseada nos descritores SIFT, HueSIFT e STIP.

Os resultados experimentais indicam o potencial do método proposto para a tarefa de sumarização de vídeos. Análises estatísticas mostraram que os resumos produzidos pela abordagem proposta alcançaram qualidade estatisticamente equivalente aos trabalhos que obtiveram as melhores avaliações em relação a taxa de inclusão dos itens do *ground truth* na tarefa de sumarização do TRECVID 2007.

Quando comparados entre si, os resumos produzidos pelos descritores SIFT, HueSIFT e STIP apresentaram resultados estatisticamente equivalentes para as medidas avaliadas, com o HueSIFT apresentando a melhor média para a inclusão de itens do *ground truth*. Por outro lado, o resumo produzido pela união dos resumos dos descritores, obteve superioridade estatística na comparação com os resumos produzidos pelo SIFT e STIP, e equivalência quando comparado ao resumo produzido pelo HueSIFT. Essas análises geram a evidência que a informação de cor possui relevância para a sumarização de vídeos baseada em pontos de interesse e que a união dos resumos melhora a qualidade do resumo final, visto que foi o sistema que obteve as melhores médias para

as medidas avaliadas.

5.1 Trabalhos futuros

A abordagem proposta alcançou bons resultados. Porém, foram identificados alguns pontos na metodologia que ainda podem ser mais explorados. Por exemplo:

- Adaptar a metodologia proposta para outros tipos de vídeo. Tais como: noticiários, seriados, documentário e etc.
- Investigar e testar outras medidas de distância além da Euclidiana. Isso pode ajudar para uma melhor comparação entre os histogramas que representam os segmentos.
- Investigar o uso de outros algoritmos de agrupamento. Por exemplo, algoritmos que não necessitem a especificação do número de grupos, tais como os de agrupamento hierárquico.
- Definir e testar novas métricas de importância para a escolha dos segmentos mais representativos.
- Definir um método com baixo custo computacional para a detecção automática de claquetes.
- Empregar outros tipos de características, como por exemplo, informações de áudio e texto. É visto na literatura que a combinação destas características pode levar a melhores resultados.

Referências Bibliográficas

Agnihotri, L.; Dimitrova, N. & Kender, J. R. (2004). Design and evaluation of a music video summarization system. In *ICME*, pp. 1943–1946. IEEE.

Agnihotri, L.; Kender, J.; Dimitrova, N. & Zimmerman, J. (2005). Framework for personalized multimedia summarization. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, MIR '05, pp. 81--88, New York, NY, USA. ACM.

Assfalg, J.; Bertini, M.; Colombo, C.; Bimbo, A. D. & Nunziati, W. (2003). Semantic annotation of soccer videos: automatic highlights identification. *Comput. Vis. Image Underst.*, 92:285--305.

Beran, V.; Hradiš, M.; Herout, A.; Sumec, S.; Potúček, I.; Zemčik, P.; Mlích, J.; Láník, A. & ChmelaƊ, P. (2007). Video summarization at brno university of technology. In *Proceedings of the international workshop on TRECVID video summarization*, TVS '07, pp. 16--19, New York, NY, USA. ACM.

Bovik, A. C. (2009). *The Essential Guide to Video Processing*, chapter 6. Academic Press, 1 edição.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Byrne, D.; Kehoe, P.; Lee, H.; Conaire, C. O.; Smeaton, A. F.; O'Connor, N. E. & Jones, G. J. (2007). A user-centered approach to rushes summarisation via highlight-detected keyframes. In *Proceedings of the international workshop on TRECVID video summarization*, TVS '07, pp. 35--39, New York, NY, USA. ACM.

Chang, P.; Han, M. & Gong, Y. (2002). Extract highlights from baseball game video with hidden markov models. pp. 609--612.

Chen, F.; Adcock, J. & Cooper, M. (2008). A simplified approach to rushes summarization. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, TVS '08, pp. 60--64, New York, NY, USA. ACM.

Chiu, P.; Girgensohn, A.; Polak, W.; Rieffel, E. & Wilcox, L. (2000). A genetic algorithm for video segmentation and summarization. In *In IEEE International Conference on Multimedia and Expo*, pp. 1329--1332.

Christel, M. G.; Hauptmann, A. G.; Lin, W.-H.; Chen, M.-Y.; Yang, J.; Maher, B. & Baron, R. V. (2008). Exploring the utility of fast-forward surrogates for bbc rushes. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, TVS '08, pp. 35--39, New York, NY, USA. ACM.

Coldefy, F. & Bouthemy, P. (2004). Unsupervised soccer video abstraction based on pitch, dominant color and camera motion analysis. In *Proceedings of the 12th annual ACM international conference on Multimedia*, MULTIMEDIA '04, pp. 268--271, New York, NY, USA. ACM.

Cooper, M. D. & Foote, J. (2002). Summarizing video using non-negative similarity matrix factorization. In *IEEE Workshop on Multimedia Signal Processing*, pp. 25–28. IEEE Signal Processing Society.

Cormen, T. H.; Leiserson, C. E.; Rivest, R. L. & Stein, C. (2001). *Introduction to Algorithms*. The MIT Press, New York.

Dagtas, S. & Abdel-Mottaleb, M. (2004). Multimodal detection of highlights for multimedia content. *Multimedia Syst.*, 9(6):586–593.

deCampos, T.; Barnard, M.; Mikolajczyk, K.; Kittler, J.; Yan, F.; Christmas, W. & Windridge, D. (2011). An evaluation of bags-of-words and spatio-temporal shapes for action recognition. In *IEEE Workshop on Applications of Computer Vision (WACV)*, Kona, Hawaii.

Detyniecki, M. & Marsala, C. (2007). Video rushes summarization by adaptive acceleration and stacking of shots. In *Proceedings of the international workshop on TRECVID video summarization*, TVS '07, pp. 65--69, New York, NY, USA. ACM.

Dumont, E. & Mérialdo, B. (2010). Rushes video summarization and evaluation. *Multimedia Tools Appl.*, 48:51--68.

Ellouze, M.; Karray, H. & Alimi, A. M. (2008). Regim, research group on intelligent machines, tunisia, at trecvid 2008, bbc rushes summarization. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, TVS '08, pp. 105--108, New York, NY, USA. ACM.

Erol, B.; Lee, D.-S. & Hull, J. (2003). Multimodal summarization of meeting recordings. In *Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 3 (ICME '03) - Volume 03*, ICME '03, pp. 25--28, Washington, DC, USA. IEEE Computer Society.

Furini, M.; Geraci, F.; Montangero, M. & Pellegrini, M. (2010). Stimo: Still and moving video storyboard for the web scenario. *Multimedia Tools Appl.*, 46:47--69.

Gong, Y. & Liu, X. (2003). Video summarization and retrieval using singular value decomposition. *Multimedia Syst.*, 9:157--168.

Hanjalic, A.; Lagendijk, R. L.; Member, S. & Biemond, J. (1999). Automated high-level movie segmentation for advanced video-retrieval systems. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 580--588.

Hannon, J.; McCarthy, K.; Lynch, J. & Smyth, B. (2011). Personalized and automatic social summarization of events in video. In *Proceedings of the 16th international conference on Intelligent user interfaces*, IUI '11, pp. 335--338, New York, NY, USA. ACM.

Harris, C. & Stephens, M. (1988). A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pp. 147--151.

He, L.; Sanocki, E.; Gupta, A. & Grudin, J. (1999). Auto-summarization of audio-video presentations. In *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, MULTIMEDIA '99, pp. 489--498, New York, NY, USA. ACM.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley and Sons, Inc.

Jiang, Y.-G.; Ngo, C.-W. & Yang, J. (2007). Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, CIVR '07, pp. 494--501, New York, NY, USA. ACM.

Laganière, R.; Bacco, R.; Hocevar, A.; Lambert, P.; Païs, G. & Ionescu, B. E. (2008). Video summarization from spatio-temporal features. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, TVS '08, pp. 144--148, New York, NY, USA. ACM.

Laptev, I. (2005). On space-time interest points. *Int. J. Comput. Vision*, 64:107--123.

Le, D.-D. & Satoh, S. (2007). National institute of informatics, japan at trecvid 2007: Bbc rushes summarization. In *Proceedings of the international workshop on TRECVID video summarization*, TVS '07, pp. 70--73, New York, NY, USA. ACM.

Lee, J.; gook Lee, G. & yul Kim, W. (2003). Automatic video summarizing tool using mpeg-7 descriptors for personal video recorder. *IEEE Trans. on Cons. Elect*, 49:49--742.

Li, K.; Guo, L.; Faraco, C.; Zhu, D.; Deng, F.; Zhang, T.; Jiang, X.; Zhang, D.; Chen, H.; Hu, X.; Miller, S. & Liu, T. (2010). Human-centered attention models for video summarization. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI '10, pp. 27:1--27:8, New York, NY, USA. ACM.

Li, L.; Zhou, K.; Xue, G.-R.; Zha, H. & Yu, Y. (2011). Video summarization via transferrable structured learning. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pp. 287--296, New York, NY, USA. ACM.

Li, Z.; Schuster, G. M.; Katsaggelos, A. K. & G, B. (2004). Optimal video summarization with a bit budget constraint. In *Proceedings of Intl Conference on Image Processing (ICIP*.

Lie, W.-N. & Hsu, K.-C. (2008). Video summarization based on semantic feature analysis and user preference. In *Proceedings of the 2008 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (sutc 2008)*, pp. 486--491, Washington, DC, USA. IEEE Computer Society.

Lie, W.-N. & Lai, C.-M. (2004). News video summarization based on spatial and motion feature analysis. In Aizawa, K.; Nakamura, Y. & Satoh, S., editores, *PCM (2)*, volume 3332 of *Lecture Notes in Computer Science*, pp. 246–255. Springer.

Lienhart, R.; Pfeiffer, S. & Effelsberg, W. (1997). Video abstracting. *Communications of the ACM*, 40:55--62.

Liu, Y.; Liu, Y. & Zhang, Y. (2007). The hong kong polytechnic university at trecvid 2007 bbc rushes summarization. In *Proceedings of the international workshop on TRECVID video summarization*, TVS '07, pp. 50--54, New York, NY, USA. ACM.

Liu, Z.; Zavesky, E.; Shahraray, B.; Gibbon, D. & Basso, A. (2008). Brief and high-interest video summary generation: evaluating the at&t labs rushes summarizations. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, TVS '08, pp. 21--25, New York, NY, USA. ACM.

Lopes, A. P. B.; De Avila, S. E. F. & Peixoto, A. N. A. (2009). *A bag-of-features approach based on hue-SIFT descriptor for nude detection*, pp. 1552--1556. Number Eusipco.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91--110.

Lu, S.; King, I. & Lyu, M. R. (2004). Video summarization by video structure analysis and graph optimization. In *ICME*, pp. 1959–1962. IEEE.

Lu, S.; Lyu, M. R. & King, I. (2005). Semantic video summarization using mutual reinforcement principle and shot arrangement patterns. In *Proceedings of the 11th International Multimedia Modelling Conference*, MMM '05, pp. 60--67, Washington, DC, USA. IEEE Computer Society.

Lucas, B. D. & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence - Volume 2*, pp. 674--679, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lupatini, G.; Saraceno, C. & Leonardi, R. (1998). Scene break detection: A comparison. In *Proceedings of the Workshop on Research Issues in Database Engineering*, RIDE '98, pp. 34--, Washington, DC, USA. IEEE Computer Society.

Ma, Y.-F.; Lu, L.; Zhang, H.-J. & Li, M. (2002). A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*, MULTIMEDIA '02, pp. 533--542, New York, NY, USA. ACM.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. & Neyman, J., editores, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297. University of California Press.

Mei, T.; Zhu, C.-Z.; Zhou, H.-Q. & Hua, X.-S. (2005). Spatio-temporal quality assessment for home videos. In *Proceedings of the 13th annual ACM international conference on Multimedia*, MULTIMEDIA '05, pp. 439--442, New York, NY, USA. ACM.

Money, A. G. & Agius, H. (2008). Video summarisation: A conceptual framework and survey of the state of the art. *J. Vis. Comun. Image Represent.*, 19:121--143.

Muslea, I.; Minton, S. & Knoblock, C. A. (2002). Active + semi-supervised learning = robust multi-view learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pp. 435--442, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Over, P.; Smeaton, A. F. & Awad, G. (2008). The trecvid 2008 bbc rushes summarization evaluation. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, TVS '08, pp. 1--20, New York, NY, USA. ACM.

Over, P.; Smeaton, A. F. & Kelly, P. (2007). The trecvid 2007 bbc rushes summarization evaluation pilot. In *Proceedings of the international workshop on TRECVID video summarization*, TVS '07, pp. 1--15, New York, NY, USA. ACM.

Pan, C.-M.; Chuang, Y.-Y. & Hsu, W. H. (2007). Ntu trecvid-2007 fast rushes summarization system. In *Proceedings of the international workshop on TRECVID video summarization*, TVS '07, pp. 74--78, New York, NY, USA. ACM.

Peyrard, N. & Bouthemy, P. (2005). Motion-based selection of relevant video segments for video summarization. *Multimedia Tools Appl.*, 26:259--276.

Putpuek, N.; Le, D.-D.; Cooharojananone, N.; Satoh, S. & Lursinsap, C. (2008). Rushes summarization using different redundancy elimination approaches. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, TVS '08, pp. 100--104, New York, NY, USA. ACM.

Reiko, K. M.; Miura, K.; Hamada, R.; Ide, I. & Sakai, S. (2003). Motion based automatic abstraction of cooking videos.

Ren, J.; Jiang, J. & Eckes, C. (2008). Hierarchical modeling and adaptive clustering for real-time summarization of rush videos in trecvid'08. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, TVS '08, pp. 26--30, New York, NY, USA. ACM.

Ren, R.; Swamy, P. P.; Jose, J. M. & Urban, J. (2007). Attention-based video summarisation in rushes collection. In *Proceedings of the international workshop on TRECVID video summarization*, TVS '07, pp. 89--93, New York, NY, USA. ACM.

Richard O. Duda, Peter E. Hart, D. G. S. (2001). *Pattern Classification*, chapter Unsupervised Learning and Clustering, p. 654. Springer-Verlag, New York.

Sasongko, J.; Rohr, C. & Tjondronegoro, D. (2008). Efficient generation of pleasant video summaries. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, TVS '08, pp. 119--123, New York, NY, USA. ACM.

Shao, X.; Xu, C.; Maddage, N. C.; Tian, Q.; Kankanhalli, M. S. & Jin, J. S. (2006). Automatic summarization of music videos. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2:127--148.

Smith, M. A. (1997). Video skimming and characterization through the combination of image and language understanding techniques. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, CVPR '97, pp. 775--, Washington, DC, USA. IEEE Computer Society.

Song, Y. & Wang, W. (2009). Unified sports video highlight detection based on multi-feature fusion. In *Proceedings of the 2009 Third International Conference on Multimedia and Ubiquitous Engineering*, MUE '09, pp. 83--87, Washington, DC, USA. IEEE Computer Society.

Sundaram, H.; Chang, S.-F. & *undararn*hih-fu Chang, H. (2001). Condensing computable scenes using visual complexity and film syntax analysis. In *Proceedings of ICME 2001*, pp. 389--392.

Takahashi, Y.; Nitta, N. & Babaguchi, N. (2005). Video summarization for large sports video archives. In *ICME*, pp. 1170–1173. IEEE.

Taskiran, C. M.; Amir, A.; Ponceleon, D. B. & Delp, E. J. (2002). Automated video summarization using speech transcripts. In Yeung, M. M.; Li, C.-S. & Lienhart, R., editores, *Storage and Retrieval for Media Databases*, volume 4676 of *SPIE Proceedings*, pp. 371–382. SPIE.

Taskiran, C. M.; Pizlo, Z.; Amir, A.; Ponceleon, D. B. & Delp, E. J. (2006). Automated video program summarization using speech transcripts. *IEEE Transactions on Multimedia*, 8(4):775–791.

Tjondronegoro, D.; Chen, Y.-P. P. & Pham, B. (2003). Sports video summarization using highlights and play-breaks. In *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, MIR '03, pp. 201--208, New York, NY, USA. ACM.

Tjondronegoro, D.; Chen, Y.-P. P. & Pham, B. (2004). Integrating highlights for more complete sports video summarization. *IEEE MultiMedia*, 11:22--37.

Toharia, P.; Robles, O. D.; Pastor, L. & Rodriguez, A. (2008). Combining activity and temporal coherence with low-level information for summarization of video rushes. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, TVS '08, pp. 70--74, New York, NY, USA. ACM.

Truong, B. T. & Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3.

Tuytelaars, T. & Mikolajczyk, K. (2008). Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 3:177--280.

Valdés, V. & Martínez, J. M. (2007). On-line video skimming based on histogram similarity. In *Proceedings of the international workshop on TRECVID video summarization*, TVS '07, pp. 94--98, New York, NY, USA. ACM.

van de Sande, K. E. A.; Gevers, T. & Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582--1596.

Wang, J.; Xu, C.; Chng, E. & Tian, Q. (2004). Sports highlight detection from keyword sequences using hmm. In *In Proc. IEEE ICME*, pp. 27--30.

Wang, T.; Gao, Y.; Li, J.; Wang, P. P.; Tong, X.; Hu, W.; Zhang, Y. & Li, J. (2007). Thu-icrc at rush summarization of trecvid 2007. In *Proceedings of the international workshop on TRECVID video summarization*, TVS '07, pp. 79--83, New York, NY, USA. ACM.

Wu, J.-K. K.; Hong, D.; Kankanhalli, M. S. & Lim, J.-H. (2000). *Perspectives on Content-Based Multimedia Systems*. Kluwer Academic Publishers, Norwell, MA, USA.

Xiong, Z.; Radhakrishnan, R. & Divakaran, A. (2003). Generation of sports highlights using motion activity in combination with a common audio feature extraction framework. In *ICIP (1)*, pp. 5–8.

Yamasaki, K.; Shinoda, K. & Furui, S. (2008). Automatically estimating number of scenes for rushes summarization. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, TVS '08, pp. 129--133, New York, NY, USA. ACM.

Yan, J.; ; Yan, J. & Pollefeys, M. (2004). Video synchronization via space-time interest point distribution. In *In Proceedings of Advanced Concepts for Intelligent Vision Systems*.

Yu, B.; Ma, W.-Y.; Nahrstedt, K. & Zhang, H.-J. (2003). Video summarization based on user log enhanced link analysis. In *Proceedings of the eleventh ACM international conference on Multimedia*, MULTIMEDIA '03, pp. 382--391, New York, NY, USA. ACM.

Yuan, J.; Wang, H.; Xiao, L.; Zheng, W.; Li, J.; Lin, F. & Zhang, B. (2007). A formal study of shot boundary detection. In *Circuit and Systems For Video Technology, 2007, IEEE Transaction on*, pp. 168--186.

Zhao, M.; Bu, J. & Chen, C. (2003). Audio and video combined for home video abstraction. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, pp. 1520--6149.