SANDRA ELIZA FONTES DE AVILA

# UMA ABORDAGEM BASEADA EM CARACTERÍSTICAS DE COR PARA A ELABORAÇÃO AUTOMÁTICA E AVALIAÇÃO SUBJETIVA DE RESUMOS ESTÁTICOS DE VÍDEOS

Belo Horizonte – MG
26 de setembro de 2008

SANDRA ELIZA FONTES DE AVILA
Orientador: Arnaldo de Albuquerque Araújo

# UMA ABORDAGEM BASEADA EM CARACTERÍSTICAS DE COR PARA A ELABORAÇÃO AUTOMÁTICA E AVALIAÇÃO SUBJETIVA DE RESUMOS ESTÁTICOS DE VÍDEOS

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

Belo Horizonte – MG
26 de setembro de 2008

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

# Uma Abordagem Baseada em Características de Cor para a Elaboração Automática e Avaliação Subjetiva de Resumos Estáticos de Vídeos

SANDRA ELIZA FONTES DE AVILA

Dissertação defendida e aprovada pela banca examinadora constituída por:

Prof. Dr. Arnaldo de Albuquerque Araújo – Orientador
Universidade Federal de Minas Gerais

Prof. Dr. Mário Fernando Montenegro Campos
Universidade Federal de Minas Gerais

Prof. Dr. Clodoveu Augusto Davis Júnior
Universidade Federal de Minas Gerais

Prof Dr. Silvio Jamil Ferzoli Guimarães
Pontifícia Universidade Católica de Minas Gerais

Belo Horizonte – MG, 26 de setembro de 2008

*A minha mãe, Maria Eliza, aos meus irmãos, Diego e Vinicius,*
*ao meu amor, Bruno Eduardo e a toda minha família.*

# Agradecimentos

A Deus, por ter me dado saúde e coragem para enfrentar este desafio.

Ao Prof. Arnaldo de Albuquerque Araújo pela orientação e incentivo recebido durante o desenvolvimento deste trabalho.

Aos meus "co-orientadores", Ana Paula Brandão Lopes e Antonio da Luz Júnior, pelas discussões, sugestões, críticas e idéias fundamentais que colaboraram no desenvolvimento deste trabalho. Meus sinceros agradecimentos.

Aos membros da banca, Clodoveu Augusto Davis Júnior, Mário Fernando Montenegro Campos e Silvio Jamil Ferzoli Guimarães, pelas minuciosas revisões realizadas, pela boa vontade com que me receberam e pelas sugestões fornecidas.

Ao Prof. Wagner Meira Júnior pelas valiosas informações sobre a etapa de agrupamento aplicada neste trabalho.

A minha mãe, Maria Eliza, pelo amor, carinho, apoio e incentivo dado em todos os momentos de minha vida, pela grandeza de mulher que representa e principalmente pela enorme paciência com os meus momentos de "stressadinha".

Ao meu pai, Antônio Ernane, que apesar de não está mais presente, sempre foi um exemplo de pessoa dedicada e estudiosa em quem eu pude me espelhar.

Ao meu namorado, Bruno Eduardo, pelo amor, carinho, companheirismo, amizade e pela paciência necessária com os caprichos deste mestrado.

Aos meus irmãozinhos, Diego e Vinicius, pelo carinho, apoio e pelas "briguinhas" de irmãos, pois mesmo de longe eu conseguia chatear eles.

Aos colegas, Daniel Pacheco de Queiroz, que desenvolveu a página para a avaliação e a criação de resumos, e Rodrigo Silva Oliveira, que implementou o algoritmo para realizar a comparação entre os resumos.

Aos amigos do Núcleo de Processamento Digital de Imagens (NPDI/DCC/UFMG), em especial Ana Paula, Antônio, Júlia, Camillo, Natália, David, Carol, Rodrigo, Thatyene e Marcelo, pelas sugestões, conversas, momentos de risadas e descontração.

Aos amigos conquistados em BH, em especial Rodrigo, Dayane, Júlia, Ana Paula e sua linda família, Camillo, Mauricio, Thiago e Hasan.

Aos queridos usuários (faço questão de listar todos os nomes) que se dispuseram a avaliar e a criar os resumos dos vídeos: Alexandre Wagner Chagas Faria, Ana Paula Brandão Lopes, Ana Paula Lobato Ribeiro, Anderson Nunes Alves Peixoto, André Luis Meneses Silva, André Pontes Melo, Ayala Nakashima Barbosa, Bárbara Camila de Santana Carvalho, Bernardo

Ávila Pires, Bruno Araújo, Bruno Eduardo Nascimento Oliveira, Bruno Fagundes Saldanha, Carolina Andrade Silva Bigonha, Cíntia Ribeiro Andrade, Claudiane Fonseca Rodrigues, Daniel da Silva Diogo Lara, Daniel Hasan Dalip, Daniel Pacheco de Queiroz, David Lunardi Flam, Djim de Andrade Martins, Douglas Eduardo Valente Pires, Eduardo Diego Fonseca da Silva, Eduardo Santos Anton Samaan, Fabiano Muniz Belem, Felipe de Oliveira Lima, Fernando Alvarenga Drumond, Francisco Rafael de Assis Eduardo Fonseca, Gustavo de Oliveira Correa, Gustavo Lele Frizzone, Henrique Bruno Mariano Pinto, Janaina Sant'Anna Gomide, José Carlos Serufo Filho, José Samuel Chapermann, Júlia Epischina Engrácia Oliveira, Juliana Corrêa da Motta, Júlio Vitor Rodrigues de Castro, Leonardo Yukio Yoshiwara, Luam Catão Totti, Lucas Tadeu Farias de Ávila, Luciana Lorena Rodrigues, Luis Felipe Duarte Flores, Marcelo de Miranda Coelho, Marcos Vinicius Fontes de Avila, Mauricio Fernández, Mirlaine Aparecida Crepalde, Natália Cosse Batista, Nilson Santos Figueiredo Júnior, Paulo Roberto Lafeta Ferreira, Rafael Cunha de Almeida, Raphael Pena Cavalcanti, Renata Russar de Mendonca, Renato José Santos Maciel, Rodrigo Rodrigues de Carvalho, Rodrigo Silva Oliveira, Sabrina de Azevedo Silveira, Samuel Lino de Macedo, Stephanie Santos Moura Costa, Thatyene Louise Alves de Souza Ramos, Thiago Cunha de Moura Salles, Thiago Nunes Coelho Cardoso, Thiago Rocha Silva, Vinicius Almeida Castro.

Aos professores do DCC, Mariza Andrade da Silva Bigonha e Roberto da Silva Bigonha, que me receberam em Belo Horizonte com sua hospitalidade indescritível.

Aos demais professores do mestrado, em especial Mario Fernando Montenegro Campos e Jussara Marques de Almeida, pelas contribuições indiretas.

Aos funcionários do DCC, em especial Túlia, Renata e Sheila, por me ajudarem constantemente.

A minha grande família pelo afeto e pela compreensão dos momentos ausentes.

A família do meu namorado, Rose, Dijalma, Sheilinha e Débora, por torcer com bastante força pelo sucesso e pela conclusão deste trabalho.

Aos amigos do Departamento de Computação (DCOMP) da Universidade Federal de Sergipe (UFS) que vibraram com mais uma conquista.

A todas as pessoas que rezaram por mim, em especial minha vovó Flor.

Ao CNPq e a CAPES pelo imprescindível apoio financeiro.

Por fim, a todos que colaboraram direta ou indiretamente na execução deste trabalho. Muito Obrigada!!!

# Resumo

Avanços em técnicas de compressão, diminuição no custo de armazenamento e transmissões em grande velocidade, têm facilitado a forma como os vídeos são criados, armazenados e distribuídos. Devido ao aumento na quantidade de dados dos vídeos distribuídos e usados em diversas aplicações, tais como máquinas de busca e bibliotecas digitais, estes se destacam como um tipo de dado multimídia, introduzindo, porém, a necessidade de um gerenciamento mais eficiente destes dados. Tudo isto tem aberto o caminho para novas áreas de pesquisa, tal como a *sumarização automática de vídeos*. Basicamente, esta área consiste em gerar pequenos resumos de vídeos, que são classificados em dois tipos: *resumo estático* (conjunto de *quadros-chave*) ou *resumo dinâmico* (conjunto de segmentos do vídeo). Neste trabalho, é apresentada uma metodologia para a elaboração automática de resumos estáticos de vídeos. O método é baseado na extração das características de cor dos quadros do vídeo e na classificação não-supervisionada destes. Os resumos produzidos são avaliados através de usuários e comparados com abordagens da literatura pesquisada. A solução proposta apresentou, com um nível de confiança de 98%, resultados com qualidade superior em relação às abordagens comparadas.

# Abstract

Advances in compression techniques, in decreasing cost of storage, and in high-speed transmission have facilitated the way videos are created, stored and distributed. The increase in the amount of video data deployed and used in many applications, such as search engines and digital libraries, reveals not only the importance as multimedia data type, but also leads to the requirement of efficient management of video data. This management paved the way for new research areas, such as *video summarization*. Essentially, this research area consists of automatic generating a short summary of a video, which can either be a *static summary* (key-frames set) or *dynamic summary* (set of video segments). This work presents a methodology for the development of static summaries. The method is based on color feature extraction from video frames and unsupervised classification. The video summaries produced are evaluated by users and compared with approaches found in the literature. With a confidence level of 98%, the proposed solution provided results with superior quality in relation to the approaches compared.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

Os avanços em técnicas de compressão, a diminuição no custo de equipamentos para aquisição e armazenamento de vídeo e, ainda, a disponibilidade de meios de transmissão de dados em alta velocidade, têm facilitado a forma como os vídeos são criados, armazenados e distribuídos. Como conseqüência, os vídeos passaram a ser utilizados em várias aplicações práticas em sistemas como máquinas de busca, bibliotecas digitais e sistemas de segurança. Devido ao aumento na quantidade de vídeos distribuídos e utilizados em aplicações atuais, a pesquisa e o desenvolvimento de novas tecnologias são necessários para um gerenciamento mais eficiente destes dados. Entre as diversas áreas possíveis de pesquisa, a sumarização automática de vídeos é uma etapa essencial para inúmeras aplicações de vídeos, tais como indexação, navegação e recuperação por conteúdo (Truong e Venkatesh, 2007).

*Sumarização de vídeo* é o processo de extração de um resumo do conteúdo original do vídeo, cujo objetivo é fornecer rapidamente uma informação concisa do seu conteúdo, preservando a mensagem original do vídeo (Pfeiffer et al., 1996). Recentemente, o resumo automático de vídeos tem atraído o interesse dos pesquisadores devido a sua importância em diversas aplicações práticas. Como conseqüência, novos modelos e algoritmos têm sido propostos na literatura da área.

Segundo Truong e Venkatesh (2007), os tipos de resumos gerados a partir das técnicas de sumarização automática de vídeos, podem ser classificados em duas categorias principais: *keyframes* ou *video skim*. A primeira categoria, também conhecida como *representative frames*, *still-image abstracts* ou *static storyboard*, consiste na extração de um conjunto de quadros-chave[1] do vídeo original, resultando em resumos estáticos. Já a segunda categoria, também conhecida como *dynamic summary*, *moving-image abstract*, *moving storyboard* ou *summary sequence*, coleta um conjunto de tomadas[2] por meio da análise da similaridade ou da relação temporal entre os quadros, resultando em resumos dinâmicos. Uma vantagem dos resumos dinâmicos em relação aos resumos estáticos é a possibilidade de incluir elementos de áudio e movimentos, realçando, assim, tanto a expressividade quanto a informação presente no resumo. Além disso, segundo Li et al. (2001), geralmente o resumo dinâmico desperta mais

[1] Um *quadro-chave* é um quadro (imagem) que representa o conteúdo de uma certa parte do vídeo.

[2] Uma *tomada* pode ser definida como uma seqüência de quadros que apresenta uma ação contínua no tempo e no espaço que foi capturada por uma única câmera.

interesse no usuário. Por outro lado, os resumos estáticos são mais apropriados para indexação e navegação. Ademais, estes últimos permitem acessar o conteúdo do vídeo de forma não-linear, pois uma vez que os quadros-chave tenham sido extraídos, existem diversas maneiras de organizá-los além da seqüência restrita visualizada nos resumos dinâmicos, como demonstrado em (Tonomura et al., 1993; Yeung e Leo, 1997; Uchihashi et al., 1999; Girgensohn et al., 2001; Girgensohn, 2003; Ćalić et al., 2007; Wang et al., 2007). Estas maneiras permitem ao usuário compreender o conteúdo do vídeo de forma mais rápida.

Apesar dos resumos dinâmicos e estáticos possuírem características diferentes, esses podem ser obtidos de modo semelhante, sendo que a partir da obtenção de um, pode-se realizar a identificação do outro. Resumos dinâmicos podem ser criados a partir de quadros-chave através da junção de segmentos de tamanho fixo, subtomadas ou da tomada a qual o quadro-chave pertence, como apresentado em (Wu et al., 2000; Lee e Hayes, 2004). E um resumo estático pode ser obtido através da seleção de quadros-chave de cada fragmento do resumo dinâmico (Hanjalic e Zhang, 1999). Neste trabalho, o método proposto para a sumarização de vídeo está voltado para a produção de resumos estáticos.

Na literatura especializada, diferentes técnicas para gerar resumos estáticos têm sido propostas (Zhuang et al., 1998; Hanjalic e Zhang, 1999; Gong e Liu, 2000; Hadi et al., 2006; Mundur et al., 2006; Chang e Chen, 2007; Furini et al., 2007), sendo que a maioria delas baseiam-se em técnicas de agrupamento (*clustering*). Para esta técnica, a idéia é produzir resumos por meio do agrupamento de quadros/tomadas similares e apresentar um número limitado de quadros por grupo (na maioria dos casos, é selecionado um quadro por grupo). Nesta abordagem, é importante selecionar o tipo das características que serão utilizadas para representar os quadros (por exemplo, distribuição de cores, vetores de movimento, textura, forma) e medir a similaridade entre eles.

Apesar das técnicas existentes produzirem resumos com qualidade aceitável, elas geralmente utilizam técnicas de agrupamento complicadas que são computacionalmente caras e requerem um alto consumo de tempo (Furini et al., 2008). Por exemplo, em (Mundur et al., 2006) o tempo necessário para a produção de um resumo leva cerca de 10 vezes a duração do vídeo. De fato, não é aceitável que um usuário tenha que esperar 20 minutos para ter uma representação concisa de um vídeo que ele poderia ter assistido em apenas dois minutos.

Neste trabalho, é proposta uma abordagem simples, eficaz e eficiente para a sumarização automática de vídeos. O método é baseado na extração das características de cor dos quadros do vídeo e na classificação não-supervisionada destes. Os resumos produzidos foram avaliados por meio de usuários e comparados com abordagens encontradas na literatura.

## 1.1 Definição do Problema

A necessidade da tarefa de resumir conteúdos de vídeos surgiu a partir da verificação de dois fatos importantes: a) popularização da distribuição de vídeos através da Internet, e; b) a característica seqüencial do vídeo que obriga o usuário a assistir a todo o conteúdo do vídeo, ou uma boa parte dele, para descobrir do que realmente se trata. Em contraste, um conjunto

de imagens pode ser varrido aleatoriamente em busca da informação desejada. Estes fatores se tornam mais impactantes se analisados em conjunto, ou seja, um usuário se vê forçado a efetuar o *download* de todo um arquivo de vídeo, em sua maioria significativamente grande, e assistir o seu conteúdo para conseguir detectar se o vídeo é relacionado a um dado assunto desejado. O custo de tempo necessário para realização desse processo é elevado, ainda mais se observado que ao final o vídeo pode não conter o que era esperado.

Neste cenário, a elaboração de métodos capazes de possibilitar a construção automática de resumo do conteúdo de vídeos se caracteriza como uma possível solução para reduzir o custo desse processo. Visto que antes de efetuar o *download* do arquivo de vídeo propriamente dito, o usuário passará a ter como opção acessar um resumo dos principais fatos presentes no vídeo e assim decidir se é ou não de seu interesse.

O problema deste trabalho pode ser resumido na seguinte frase:

> *Dado um vídeo digital, selecionar, de forma eficiente, um conjunto conciso de quadros que seja representativo.*

## 1.2 Motivação

O presente trabalho é motivado devido ao crescimento da geração e distribuição de vídeos digitais, principalmente na internet. Com o aumento no volume de dados, cresce também a necessidade de desenvolvimento de técnicas, para possibilitar o acesso rápido às informações contidas nos vídeos, bem como obter de modo ágil uma noção dos principais objetos e eventos presentes. Resumos gerados de forma concisa e inteligente facilitarão o acesso do usuário a um grande volume substancial de vídeos de maneira eficiente e eficaz.

## 1.3 Objetivos

### 1.3.1 Objetivo Geral

Desenvolver uma abordagem eficaz e eficiente para sumarização automática de vídeos, que gere resumos estáticos concisos e representativos, possibilitando a identificação dos principais eventos presentes no vídeo.

### 1.3.2 Objetivos Específicos

1. Definir uma metodologia simples para a sumarização de vídeos que permita produzir resumos de forma eficaz e eficiente;

2. Identificar as características de cor mais adequadas para o processo de sumarização proposto;

3. Definir um método que determine o conjunto conciso de quadros-chave para a obtenção de resumos representativos;

4. Definir um método de avaliação que permita quantificar a qualidade dos resumos gerados e reduzir a subjetividade presente na tarefa de avaliação de resumos.

## 1.4 Restrições

O método de sumarização de vídeos apresentado neste trabalho foi concebido para efetuar a sumarização em qualquer tipo de vídeo (diversos gêneros e diferentes tempos de duração – vídeos curtos e vídeos longos). Todavia, as validações realizadas no presente trabalho contemplaram vídeos curtos e alguns gêneros. Esta limitação ocorreu devido à necessidade de comparar os resumos produzidos por meio da metodologia proposta com os resumos gerados pelas abordagens encontradas na literatura. Diferentemente da maioria das abordagens, que não informa os resumos gerados e tampouco permite a re-implementação do algoritmo desenvolvido, pois geralmente muitas informações são omitidas, (Mundur et al., 2006; Furini et al., 2007) disponibilizaram os seus resumos. As duas abordagens utilizaram os mesmos vídeos, que também foram utilizados neste trabalho.

Outra restrição do trabalho é a aplicação da metodologia proposta em vídeos coloridos. Como os resumos são gerados utilizando a informação de cor presente no vídeo, é necessário vídeos com cores.

## 1.5 Material

Este trabalho fez uso do repositório de vídeos Open Video[3], composto por aproximadamente 4000 vídeos, que estão classificados de acordo com o formato, o gênero, o tempo total de duração, a cor e o áudio. Para a realização dos experimentos, foram utilizados 70 vídeos deste repositório, sendo que 20 vídeos foram utilizados para fazer uma avaliação preliminar da metodologia proposta. A partir dessa avaliação, algumas mudanças foram realizadas com o objetivo de solucionar os problemas identificados na metodologia proposta inicialmente. Os 50 vídeos restantes foram usados para avaliar o desempenho da nova metodologia proposta para o problema da sumarização automática de vídeos.

## 1.6 Contribuições

A principal contribuição deste trabalho é a definição de um método simples, eficaz e eficiente para a elaboração automática de resumos estáticos de vídeos. A metodologia proposta integra as vantagens dos conceitos apresentados nos trabalhos de referência na área de sumarização de vídeos de modo a aproveitar em um único método as contribuições relevantes de técnicas anteriormente propostas, contribuindo para o aprimoramento do estado da arte desta área de aplicação.

[3] http://www.open-video.org

Outra contribuição importante é a definição de um método de avaliação de resumos que reduz a subjetividade presente nesta tarefa e permite, principalmente, realizar de forma rápida comparações entre diferentes abordagens.

Também, pode ser citada como contribuição a investigação experimental realizada neste trabalho. Todas conclusões obtidas foram feitas com embasamento estatístico, que confirmou a superioridade da metodologia proposta em relação às abordagens encontradas na literatura.

## 1.7 Trabalhos Relacionados

Diversas abordagens foram estudadas para propor a solução do presente trabalho. A seguir, são relatadas as abordagens que estão relacionadas com a solução proposta.

Zhuang et al. (1998) propõem uma técnica de agrupamento não-supervisionada para a produção de resumos. O vídeo é segmentado em tomadas e os grupos são obtidos utilizando as características extraídas através do histograma de cor, para o espaço de cor HSV. O algoritmo utiliza um limiar $\delta$ para controlar a densidade da classificação. Antes de um quadro ser classificado em um determinado grupo, a similaridade entre o quadro e o centróide é computada. Caso este valor seja menor que $\delta$, então o quadro não está próximo o suficiente para ser inserido neste grupo. A seleção dos quadros-chave é realizada apenas nos grupos nos quais os seus tamanhos são maiores que $N/k$, onde $N$ representa o número de quadros do vídeo e $k$ o número de grupos. Para cada grupo-chave, o quadro mais próximo do centróide é selecionado como quadro-chave. Segundo os autores, o método proposto é eficiente e eficaz. Entretanto, não são realizadas avaliações do método e dos resumos gerados que possam validar estas afirmações.

Hanjalic e Zhang (1999) desenvolvem um procedimento automático para sumarização de vídeo através da validação da análise de grupos (Jain et al., 1999). O método proposto está dividido em três etapas. Primeiro, um algoritmo particional de agrupamento é aplicado $n$ vezes para todos os quadros do vídeo, os quais são representados através do histograma de cor para o espaço YUV. O número de grupos inicialmente é igual a um e é incrementado de um em um a cada aplicação do algoritmo. Desta maneira, diferentes possibilidades de grupos são obtidas. Na segunda etapa, a combinação ótima dos grupos é determinada automaticamente através da validação da análise de grupos. E por último, para cada grupo, o quadro mais próximo do centróide representa o quadro-chave. O algoritmo é executado para segmentos de filmes (não especificados). Os autores avaliam a análise de grupos proposta, mas não fazem a avaliação dos resumos produzidos.

Gong e Liu (2000) propõem um técnica para sumarização de vídeo baseada na decomposição em valores singulares (do inglês *Singular Value Decomposition* – SVD). Inicialmente, o vídeo é segmentado em quadros. Para reduzir o número de quadros a serem analisados, os autores utilizam uma subamostragem dos quadros (um quadro a cada 10 quadros). A extração de características destes é realizada através do histograma de cor para o espaço RGB. Cada quadro é divido em blocos de dimensão $3 \times 3$ e para cada bloco é calculado o histograma, gerando assim nove histogramas para cada quadro. Para reduzir a dimensão do vetor de ca-

racterísticas é aplicado o algoritmo SVD. A partir dos vetores redimensionados, as distâncias entre os vetores são calculadas. O menor valor representa o limiar utilizado para gerar os grupos. Para cada grupo, o quadro mais próximo do centróide representa o quadro-chave. Os testes são realizados em vídeos sobre documentários, debates políticos, noticiários e segmentos de filmes. No entanto, os resultados obtidos não são comparados com outras abordagens da literatura.

Mundur et al. (2006) apresentam um método baseado na triangulação de Delaunay para agrupar os quadros similares do vídeo. Inicialmente, o vídeo é segmentado em quadros e é feita uma subamostragem dos quadros (um quadro a cada 30 quadros). Para a extração de características é aplicado o histograma de cor no espaço HSV. Cada quadro é representado por um vetor de características com 256 dimensões e a seqüência do vídeo é representada por uma matriz $A$. Para reduzir a dimensão da matriz é aplicada a técnica de análise dos componentes principais (do inglês *Principal Component Analysis* – PCA), que gera uma matriz $B$ de dimensão $n \times d$, onde $n$ é número total de quadros selecionados (subamostragem) e $d$ é o número de componentes principais. Assim, cada quadro é representado por um vetor $d$-dimensional e o diagrama de Delaunay é construído para $n$ pontos em $d$-dimensões. Os grupos são obtidos de acordo com as arestas de separação no diagrama. A remoção destas arestas identificam os grupos no diagrama. Para cada grupo, o quadro-chave é representado pelo quadro mais próximo do centróide. Para avaliar a qualidade dos resumos três métricas são definidas: 1) fator de importância, denota a importância do conteúdo representado por cada grupo através do número de quadros do grupo, 2) fator de compressão, quantifica a relação entre o número de quadros do resumo e do vídeo e 3) fator de sobreposição, é utilizada para fazer uma comparação significativa entre os resumos gerados pela técnica proposta e os resumos do Open Video. Apesar do método proposto ser inteiramente automático (não exige que o número de grupos seja pré-definido e não utiliza limiares), ele consome muito tempo para gerar um resumo, cerca de 10 vezes a duração do vídeo. Além disso, o método não preserva a ordem temporal do vídeo.

Furini et al. (2007) propõem uma abordagem, denominada VISTO (*VIsual STOryboard*), que aplica um algoritmo de grupo para selecionar os quadros mais representativos. VISTO está divido em três etapas. Primeiro, os quadros do vídeo são descritos através do histograma de cor no espaço HSV. Cada quadro é representado por um vetor de características com 256 dimensões. Na segunda etapa, os autores utilizam uma versão melhorada do algoritmo *furthest-point-first* (Geraci et al., 2006) para obter os grupos. Para determinar o número de agrupamentos é calculada a distância par-a-par entre os quadros. Se a distância for maior que um limiar $\Gamma$, então o número de grupos é incrementado. A última etapa consiste em eliminar quadros redundantes e inexpressivos (por exemplo, um quadro todo preto). Para avaliar a qualidade dos resumos, os autores propõem uma avaliação subjetiva. Um conjunto de 20 pessoas são questionadas, de acordo com uma escala que varia entre um e cinco, se o resumo gerado representa bem o conteúdo do vídeo original. A média da opinião das pessoas indica a qualidade do resumo. Os resumos gerados pelo método proposto são comparados com os resumos gerados em (Mundur et al., 2006) e os resumos do Open Video.

De acordo com a análise das abordagens encontradas na literatura, pôde-se observar que na seleção dos quadros-chave são utilizadas diversas características visuais e estatísticas. Estes atributos podem influenciar consideravelmente no aumento do custo computacional e principalmente na qualidade do resumo resultante. Outro ponto identificado é que a extração de características pode produzir matrizes com dimensões elevadas. Conseqüentemente, técnicas matemáticas são utilizadas para tentar reduzir o tamanho da matriz, como visto em (Gong e Liu, 2000; Mundur et al., 2006), o que requer mais tempo de processamento. Uma deficiência evidente nos trabalhos analisados é a falta da avaliação dos resultados e da comparação com outras abordagens.

## 1.8 Organização deste Trabalho

Este trabalho está organizado como segue. No Capítulo 2, são descritos os aspectos teóricos necessários para o embasamento do presente trabalho. No Capítulo 3, é apresentada uma revisão teórica sobre o processo de sumarização automática de vídeos, especificamente voltada para a geração de resumos estáticos. A metodologia desenvolvida neste trabalho e os métodos de avaliação propostos para estimar a qualidade dos resumos produzidos são descritos no Capítulo 4. No Capítulo 5, é apresentada a base de dados utilizada nos experimentos, os resultados obtidos e a análise destes. As conclusões gerais e linhas de futuras pesquisas são delineadas no Capítulo 6 deste trabalho.

# Capítulo 2

# Fundamentação Teórica

Neste capítulo, são apresentados os aspectos teóricos necessários para o embasamento do presente trabalho. O capítulo está organizado da seguinte forma: na Seção 2.1, são descritos os conceitos básicos sobre vídeo digital; na Seção 2.2, são abordados os conceitos relacionados à tarefa de descrição do conteúdo visual de vídeo, principalmente a característica de cor; e na Seção 2.3, é apresentado o processo de classificação, especificamente a técnica de classificação não-supervisionada.

## 2.1 Vídeo Digital

*Vídeo digital* é uma seqüência de quadros (imagens) em formato digital que, quando reproduzidos um a um, em determinada velocidade, apresentam a ilusão de movimento. Tipicamente, são usados 24 quadros por segundo para se obter tal efeito. Um vídeo pode ainda conter um canal de áudio sincronizado à seqüência de imagens.

A taxa de quadros, medida em quadros por segundos (do inglês, *frames per second* – fps), define quantos quadros são mostrados a cada segundo do vídeo. Quanto maior a taxa de quadros, mais suave será a aparência do movimento durante a reprodução do vídeo e, como conseqüência, maior será a quantidade de quadros para armazenar. Para reduzir este problema, técnicas de compressão de vídeo, tais como MPEG–1, MPEG–2, MPEG–7 e MJPEG (Poynton, 2003; Woods, 2006), são normalmente utilizadas.

Os vídeos digitais possuem uma estrutura sintática composta de uma hierarquia de componentes, como apresentado na Figura 2.1. No nível mais baixo da hierarquia estão os *quadros*, que equivalem as imagens do vídeo. No próximo nível, os quadros são agrupados em tomadas. Uma *tomada* (do inglês, *shot*) (Beaver, 1994) é uma seqüência de quadros que apresenta uma ação contínua no tempo e no espaço que foi capturada por uma única câmera. As tomadas são agrupadas em cenas. Uma *cena* (Beaver, 1994) é normalmente composta por um pequeno número de tomadas inter-relacionadas que são unificadas por características semelhantes e pela proximidade temporal. O vídeo é formado da junção de todas as cenas.

Tipicamente, um vídeo é composto por uma grande quantidade de informação. Representações mais concisas do vídeo são úteis para auxiliar o processo de indexação ou navegação

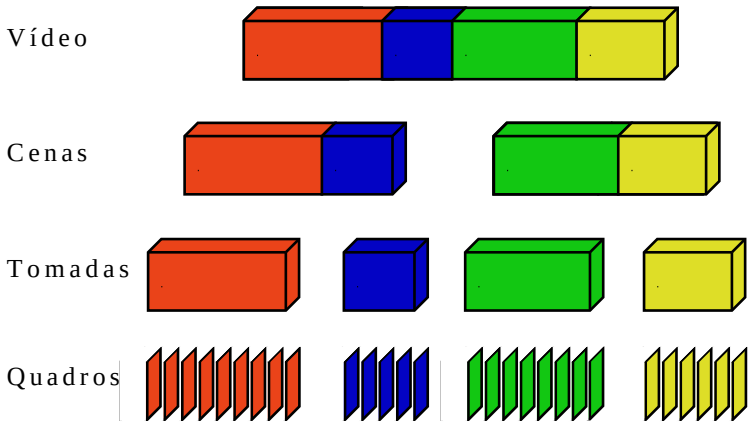



Figura 2.1: Estrutura hierárquica de um vídeo.

do vídeo, por exemplo. Uma forma de representar segmentos do vídeo é a utilização de um ou mais quadros-chave. Um *quadro-chave* (do inglês, *keyframe*) é um quadro que representa o conteúdo de uma certa tomada ou cena do vídeo. Este quadro deve ser o mais representativo possível. Considere como exemplo a Figura 2.2. Para obter um sumário visual conciso do vídeo em questão, pode-se admitir que há grande redundância entre os quadros da seqüência, de modo que um único deles poderia representar de modo satisfatório todo o conteúdo do segmento para fins de navegação, por exemplo.

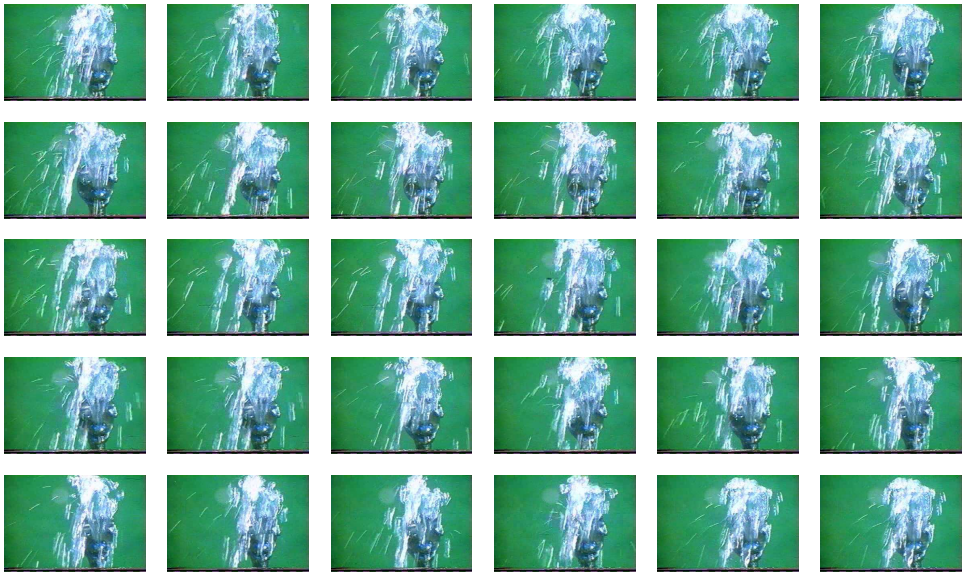


Figura 2.2: Primeiros trinta quadros do vídeo *The Great Web of Water, segment 1* (disponível via *Open Video*).

## 2.2 Descritor de Imagem

Para a geração de resumos de vídeo, é desejável descrever o seu conteúdo visual. Isto é, descrever a seqüência de imagens pelas características de seus objetos, tais como cor, forma, movimento e textura, normalmente representadas em vetores de características. Descritores de imagens são utilizados para extrair e comparar estes vetores, viabilizando a identificação do conteúdo visual em um vídeo. Isto é necessário para que o resumo produzido tenha uma relação com o conteúdo visual dos quadros.

Um *descritor de imagem* $\mathcal{D}$ (da Silva Torres e Falcão, 2006) é definido como uma tupla ($\varepsilon_{\mathcal{D}}$,$\delta_{\mathcal{D}}$), onde $\varepsilon_{\mathcal{D}} : I \rightarrow \mathbb{R}^n$ é uma função que extrai o vetor de características $\overrightarrow{v}_I$ da imagem $I$, e $\delta_{\mathcal{D}} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ é uma função de distância que computa a similaridade entre duas imagens considerando seus vetores de características. Quanto menor a distância, maior a similaridade entre as imagens.

Um *vetor de características* $\overrightarrow{v}_I$ da imagem $I$ é definido como um ponto no espaço $\mathbb{R}^n$, tal que $\overrightarrow{v}_I = (v_1, v_2, \ldots, v_n)$, onde $n$ é a dimensão do vetor.

Descritores de imagem devem possuir propriedades que garantam sua efetividade. Uma propriedade fundamental é a caracterização única de um determinado objeto a partir do vetor de características, facilitando a distinção entre imagens visualmente diferentes, de acordo com alguma métrica. Outras propriedades importantes são: geração de vetores de características compactos (que requerem pouco espaço de armazenamento) e função de extração computacionalmente eficiente.

### 2.2.1 Característica de Cor

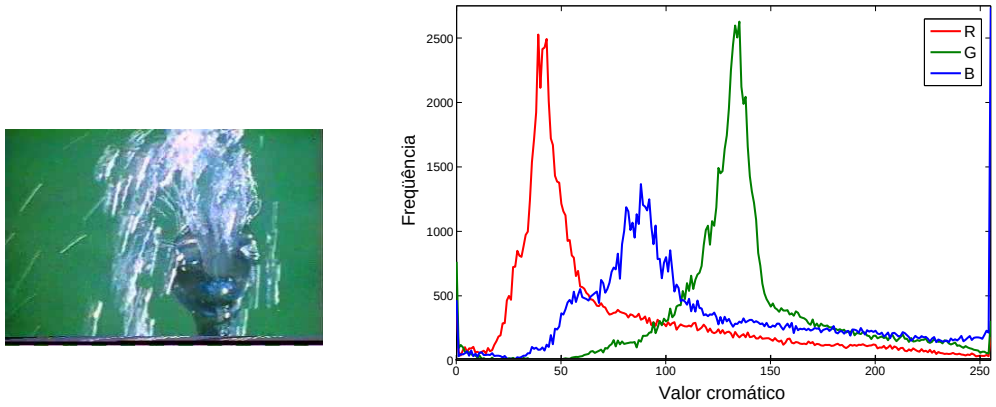
As características pertencentes a este grupo de informações são obtidas por meio da análise da distribuição dos valores de *pixels* existentes na imagem. A cor de uma imagem pode ser representada por diferentes modelos de cores. Em certas situações, antes que a extração de características possa ser realizada, uma mudança no modelo de cores da imagem, no intuito de obter uma maior agilidade no processo de extração e posterior comparação, deve ser executada. No entanto, a dinâmica do funcionamento geral dos métodos é semelhante para todos os modelos de cores existentes. A técnica histograma de cor, apresentada a seguir, é a mais simples.

#### Histograma de Cor

O *histograma de cor* (Swain e Ballard, 1991) representa a distribuição da freqüência de ocorrência dos valores cromáticos em uma imagem. Essa técnica é computacionalmente simples[1] e é robusta a pequenas mudanças do movimento da câmera. Além disso, objetos distintos freqüentemente possuem histogramas diferentes. Devido a essas características, a técnica é amplamente utilizada em sistemas de sumarização automática de vídeos (Zhuang et al., 1998; Hanjalic e Zhang, 1999; Gong e Liu, 2000; Mundur et al., 2006; Furini et al., 2007).

[1] Complexidade de execução igual a $O(n)$, onde $n$ é o número de *pixels* da imagem.

Na Figura 2.3, é ilustrada uma representação gráfica do histograma de cor gerado para uma determinada imagem no sistema RGB.

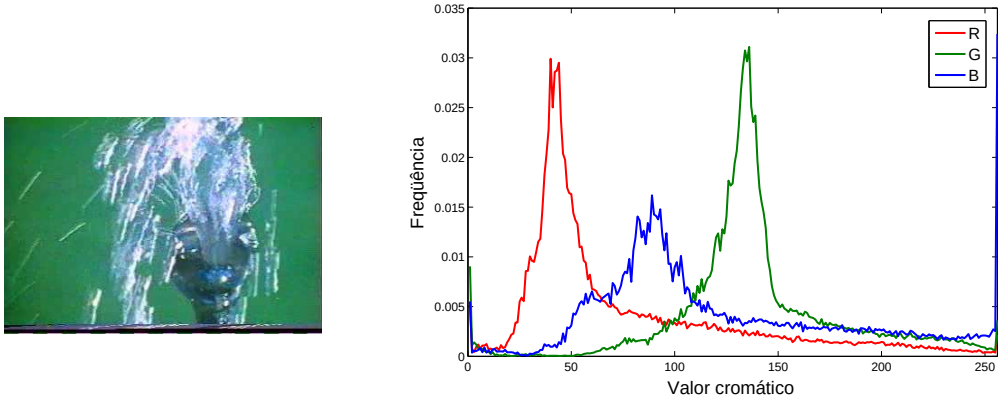



(a) Imagem (b) Histograma de cor

Figura 2.3: Ilustração do histograma de cor.

### Histograma Normalizado

O *histograma normalizado* é obtido com a transformação dos valores de um histograma para uma faixa de valores conhecida. Na maioria das aplicações, a faixa de valores adotada é [0,1]. A operação de normalização é realizada por meio da divisão do histograma original pelo número de *pixels* da imagem. O histograma normalizado possui o mesmo número de cores avaliadas, que o histograma original, a única diferença entre eles está nos valores de cada uma das cores. Na Figura 2.4, é ilustrada uma representação gráfica do histograma normalizado gerado para a mesma imagem da Figura 2.3.

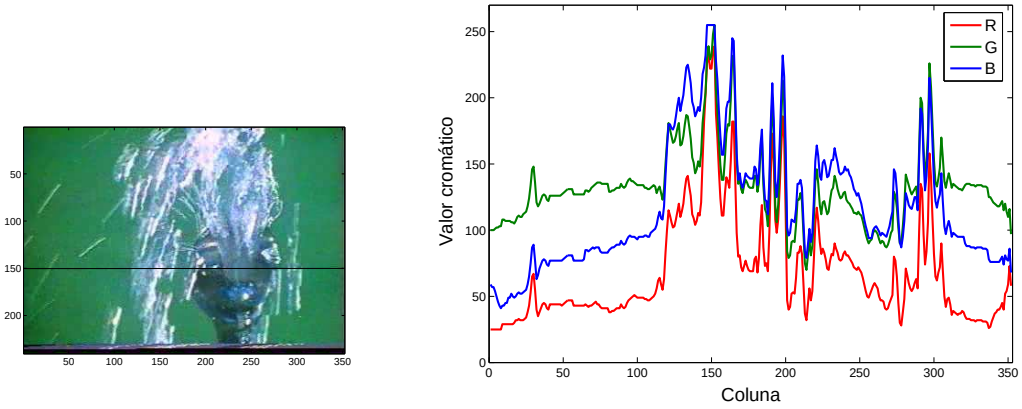



(a) Imagem (b) Histograma normalizado

Figura 2.4: Ilustração do histograma normalizado.

### Perfil de Linha

O *perfil de linha* da imagem representa os valores cromáticos presentes nesta linha. O perfil pode ser obtido em qualquer direção, sendo que as direções comumente aplicadas são: horizontal e vertical. Na Figura 2.5, é ilustrada uma representação gráfica do perfil horizontal na linha 150 da imagem.




(a) Imagem

(b) Perfil horizontal na linha 150

Figura 2.5: Ilustração do perfil horizontal.

### Modelo de Cores

Um modelo de cor é uma especificação de um sistema de coordenadas tridimensionais e um subespaço dentro deste sistema onde cada cor é representada por um único ponto. Diversos modelos de cores podem ser encontrados na literatura (Gonzalez e Woods, 2006), entretanto os modelos freqüentemente utilizados para o processamento de imagens são o RGB (*Red*, *Green*, *Blue*), usado para monitores coloridos e câmeras de vídeo em cores e o modelo HSV (*Hue*, *Saturation*, *Value*), usado para manipulação de imagens coloridas (Gonzalez e Woods, 2006). Estes dois modelos são abordados a seguir.

### Modelo RGB

No modelo RGB, cada cor é decomposta nos seus componentes espectrais primários de vermelho ($R$), verde ($G$) e azul ($B$). Este modelo baseia-se em um sistema de coordenadas cartesianas. O subespaço de cores de interesse é o cubo mostrado na Figura 2.6, no qual os valores RGB estão nos três cantos; ciano, magenta e amarelo estão nos outros três cantos, preto está na origem; e branco está no canto mais distante da origem. Nesse modelo, a escala de cinza estende-se do preto ao branco ao longo da linha diagonal que conecta estes dois pontos, e as cores são pontos sobre ou dentro do cubo, definidas por vetores estendendo-se a partir da origem. Por conveniência, assume-se que todos os valores de cor foram normalizados, de modo que o cubo é unitário, isto é, todos os valores de $R$, $G$ e $B$ são assumidos estar no

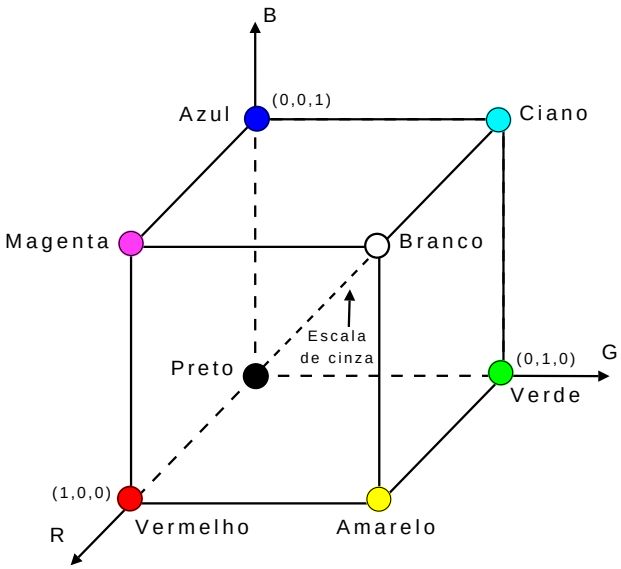



Figura 2.6: Cubo de cores RGB (Gonzalez e Woods, 2006).

intervalo [0,1] (Gonzalez e Woods, 2006).

O modelo RGB não é capaz de representar todas as cores perceptíveis pelo sistema visual humano, mas é uma boa aproximação. Uma variação de luminosidade ou saturação no modelo RGB varia de forma irregular dentro do espaço de cor, sendo necessárias outras formas de representação para capturar as informações de cor mais de acordo com o sistema visual humano.

### Modelo HSV

O modelo HSV descreve as cores de forma mais intuitiva que o modelo RGB (Foley et al., 1990). Os parâmetros de cor utilizados nesse modelo são o matiz, a saturação e a luminância. *Matiz* ($H$) é um atributo que descreve uma cor pura (amarelo puro, laranja ou vermelho), enquanto *saturação* ($S$) dá uma medida do grau de diluição de uma cor pura por uma luz branca. Cores tais como cor-de-rosa (vermelho e branco) e lilás (violeta e branco) são menos saturadas, com o grau de saturação sendo inversamente proporcional à quantidade de luz branca adicionada. A *luminância* ($V$), ou valor, é a intensidade da luz refletida (claro/escuro) (Gonzalez e Woods, 2006).

A representação do modelo HSV é ilustrado na Figura 2.7. O matiz, que corresponde às arestas ao redor do eixo vertical, varia de 0° (vermelho) a 360°, e o ângulo entre os vértices é de 60°. A saturação varia de 0 a 1 e é representada como sendo a razão entre a pureza de um determinado matiz e a sua pureza máxima ($S = 1$). Um determinado matiz possui $\frac{1}{4}$ de pureza em $S = 0{,}25$. Quando $S = 0$ tem-se a escala de cinzas. A luminância varia de 0 (no vértice do cone), que representa a cor preta, a 1 (na base), onde as intensidades das cores são máximas.

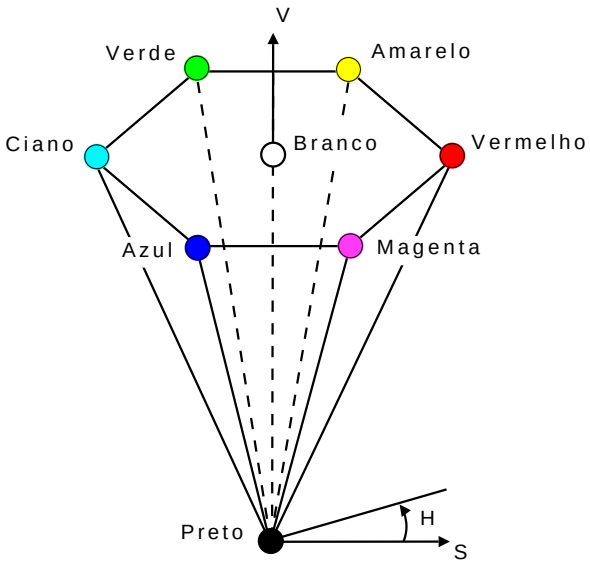



Figura 2.7: Cone hexagonal HSV (Foley et al., 1990).

De acordo com Gonzalez e Woods (2006), esse modelo deve sua utilidade a dois fatos principais. Primeiramente, o componente de intensidade $V$ é desacoplado da informação de cor na imagem. A importância deste desacomplamento é que o componente de intensidade pode ser processado sem afetar o seu conteúdo de cor. Em segundo lugar, os componente de matiz e saturação são intimamente relacionados à percepção humana de cores. Essas características tornam o modelo adequado para o desenvolvimento de algoritmos de processamento de imagens baseados em propriedades do sistema visual humano.

### Conversão de RGB para HSV

A conversão das cores no modelo RGB para o modelo HSV é dada pelas seguintes equações, descritas por Foley et al. (1990).

$$
\begin{aligned}
V &= \max \\
S &= \frac{\max - \min}{\max} \\
H &= \begin{cases}
\dfrac{G-B}{\max - \min} \times 60 & \text{se} \quad \max = R \\
2 + \dfrac{B-R}{\max - \min} \times 60 & \text{se} \quad \max = G \\
4 + \dfrac{R-G}{\max - \min} \times 60 & \text{se} \quad \max = B
\end{cases}
\end{aligned}
$$

onde $max$ representa o valor máximo entre os valores $R$, $G$ e $B$ e $min$ representa o valor mínimo no pixel.

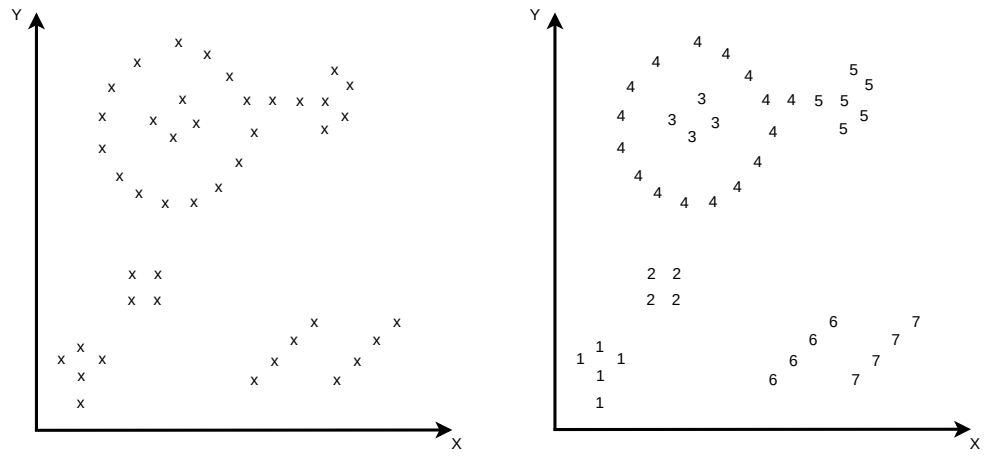
## 2.3 Classificação

*Classificação* é uma das mais freqüentes tarefas de tomada de decisão da atividade humana. Um problema de classificação acontece quando um objeto precisa ser associado a um determinado grupo ou classe baseando-se em um número de características observadas e relacionadas àquele objeto. Os métodos de classificação estão divididos em: supervisionados e não-supervisionados.

O processo de classificação *supervisionado* é caracterizado pela existência de um conjunto de padrões do qual se conhece, previamente, as classes em que estes estão divididos e em qual dessas, pelo menos um dos padrões está inserido. Ou seja, são fornecidos padrões rotulados e o problema consiste em rotular novos padrões, ainda não-rotulados. Tipicamente, os padrões rotulados são utilizados para aprender a descrição da classe (treinamento) e esta informação aprendida, por sua vez, é usada para rotular um novo padrão.

Diferentemente, um método de classificação é considerado como *não-supervisionado* quando trabalha com um conjunto de padrões dos quais não se conhecem as informações das classes existentes. Nesse tipo de classificação, também denominado de agrupamento (*clustering*), o problema consiste em agrupar um conjunto de padrões não-rotulados, usualmente representados na forma de vetores de características ou pontos em um espaço multidimensional, em grupos (*clusters*) de acordo com alguma medida de similaridade. Sendo assim, uma vez definidos os grupos, os padrões também estarão "rotulados", mas o rótulo neste caso é ditado pelos próprios padrões que compõem cada grupo. Nesta seção, será abordada apenas a técnica de agrupamento.

Na Figura 2.8, é apresentado um exemplo ilustrativo de agrupamento. Na parte (a), é apresentado um conjunto de padrões, onde cada elemento é representado pelo símbolo 'x'. Na parte (b), é apresentado o resultado do agrupamento aplicado no conjunto, com cada elemento 'x' sendo rotulado com um número identificador do conjunto a que pertence ao final do agrupamento.




(a) (b)

Figura 2.8: Ilustração da técnica de agrupamento (Jain et al., 1999).

Segundo Jain e Dubes (1988), a aplicação de técnicas de agrupamento envolve as etapas descritas a seguir (veja Figura 2.9).

**Representação dos padrões** Consiste na definição do número, tipo e modo de apresentação das características que descrevem cada padrão. A representação pode ser realizada por meio da extração ou seleção de características. A *extração de características* aplica uma ou mais transformações nos padrões de entrada de modo a salientar uma ou mais características dentre aquelas que estão presentes nos padrões. A *seleção de características* visa identificar o subconjunto mais efetivo das características disponíveis para descrever cada padrão.

**Definição da medida de similaridade** Função de distância ou medida de similaridade aplicada entre pares de padrões para obter o grau de similaridade entre eles. Algumas medidas de similaridade são exemplificadas na Seção 2.3.1.

**Desenvolvimento ou seleção do algoritmo de agrupamento** O agrupamento pode ser executado de várias formas. Os grupos podem ser definidos como conjuntos *hard* (um padrão pertence ou não-pertence a um dado grupo) ou *fuzzy* (um padrão pode apresentar graus de pertinência aos grupos). O processo de agrupamento, geralmente, é classificado como *hierárquico* (processo recursivo de junções ou separações de grupos baseado na similaridade entre os padrões) ou *particional* (identificação das partições por meio de otimização de uma função de custo). Entretanto, além dessas classificações, os métodos de agrupamentos podem ser baseados em densidade, teoria dos grafos, pesquisa combinatória, redes neuronais, entre outros. Dentre os diversos algoritmos de agrupamento, o algoritmo particional *k-means* (MacQueen, 1967) é um dos mais aplicados. Este é descrito na Seção 2.3.2.

**Avaliação do resultado** Basicamente, consiste em avaliar a qualidade dos grupos obtidos pela técnica de agrupamento aplicada. Um problema levantado nesta etapa é que, geralmente a análise do agrupamento baseia-se em algum critério de otimalidade definido de forma subjetiva.

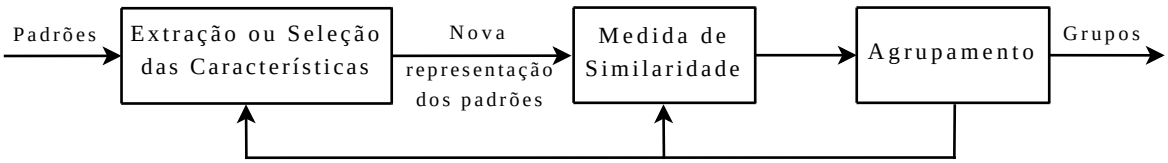



Figura 2.9: Etapas de um processo de agrupamento.

### 2.3.1 Medidas de Similaridade

O conceito de similaridade é fundamental para o processo de agrupamento, pois, se dois padrões são similares de acordo com algum critério, então serão agrupados em um mesmo

grupo, caso contrário, serão agrupados em grupos diferentes. Para isto, a definição de medidas, ou métricas, de distância que permitam comparar padrões pertencentes a um mesmo espaço de características é essencial para a maioria dos processos de agrupamento. Quanto menor for a distância entre um par de padrões maior é a similaridade entre eles.

Uma *métrica de distância* é uma função $d(.,.)$ que satisfaz as seguintes propriedades:

(i) Simetria: $d(\overrightarrow{v}_{I_1}, \overrightarrow{v}_{I_2}) = d(\overrightarrow{v}_{I_2}, \overrightarrow{v}_{I_1})$;

(ii) Positividade: $0 < d(\overrightarrow{v}_{I_1}, \overrightarrow{v}_{I_2}) < \infty$, $\overrightarrow{v}_{I_1} = \overrightarrow{v}_{I_2} \Leftrightarrow d(\overrightarrow{v}_1, \overrightarrow{v}_2) = 0$;

(iii) Desigualdade triangular: $d(\overrightarrow{v}_{I_1}, \overrightarrow{v}_{I_3}) \leq d(\overrightarrow{v}_{I_1}, \overrightarrow{v}_{I_2}) + d(\overrightarrow{v}_{I_2}, \overrightarrow{v}_{I_3})$.

onde $\overrightarrow{v}_{I_1}$, $\overrightarrow{v}_{I_2}$ e $\overrightarrow{v}_{I_3}$ representam três vetores de características das imagens $I_1$, $I_3$ e $I_3$, respectivamente.

Na Tabela 2.1, são apresentadas algumas medidas de similaridade.

Tabela 2.1: Medidas de Similaridade.

| **Medida** | **Fórmula** | |
|---|---|---|
| Distância Euclidiana (Androutsos et al., 1998) | $d(\overrightarrow{v}_{I_1}, \overrightarrow{v}_{I_2}) = \sqrt{\sum_{i=1}^{n} (\overrightarrow{v}_{I_1 i} - \overrightarrow{v}_{I_2 i})^2}$ | (2.1) |
| Distância de Manhattan (Androutsos et al., 1998) | $d(\overrightarrow{v}_{I_1}, \overrightarrow{v}_{I_2}) = \sum_{i=1}^{n} \lvert \overrightarrow{v}_{I_1 i} - \overrightarrow{v}_{I_2 i} \rvert$ | (2.2) |
| Distância de Canberra (Androutsos et al., 1998) | $d(\overrightarrow{v}_{I_1}, \overrightarrow{v}_{I_2}) = \sum_{i=1}^{n} \frac{\lvert \overrightarrow{v}_{I_1 i} - \overrightarrow{v}_{I_2 i} \rvert}{\lvert \overrightarrow{v}_{I_1 i} \rvert + \lvert \overrightarrow{v}_{I_2 i} \rvert}$ | (2.3) |
| Distância de Bray–Curtis (Androutsos et al., 1998) | $d(\overrightarrow{v}_{I_1}, \overrightarrow{v}_{I_2}) = \sum_{i=1}^{n} \frac{\lvert \overrightarrow{v}_{I_1 i} - \overrightarrow{v}_{I_2 i} \rvert}{\overrightarrow{v}_{I_1 i} + \overrightarrow{v}_{I_2 i}}$ | (2.4) |
| Distância $\mathcal{X}^2$ (Cámara-Chávez, 2007) | $d(\overrightarrow{v}_{I_1}, \overrightarrow{v}_{I_2}) = \sum_{i=1}^{n} \frac{(\overrightarrow{v}_{I_1 i} - \overrightarrow{v}_{I_2 i})^2}{\overrightarrow{v}_{I_1 i} + \overrightarrow{v}_{I_2 i}}$ | (2.5) |
| Similaridade de Cosseno (Cámara-Chávez, 2007) | $d(\overrightarrow{v}_{I_1}, \overrightarrow{v}_{I_2}) = \sum_{i=1}^{n} \frac{\overrightarrow{v}_{I_1 i} \times \overrightarrow{v}_{I_2 i}}{\lVert \overrightarrow{v}_{I_1 i} \rVert \times \lVert \overrightarrow{v}_{I_2 i} \rVert}$ | (2.6) |
| Intersecção de Histogramas (Swain e Ballard, 1991) | $d(\overrightarrow{v}_{I_1}, \overrightarrow{v}_{I_2}) = \frac{\sum_{i=1}^{n} \min(\overrightarrow{v}_{I_1 i}, \overrightarrow{v}_{I_2 i})}{\sum_{i=1}^{n} \overrightarrow{v}_{I_2 i}}$ | (2.7) |

### 2.3.2 Algoritmo *k-means*

O algoritmo *k-means* (*k*-médias) (MacQueen, 1967) é um dos mais simples algoritmos de classificação não-supervisionada e mais amplamente utilizado, apresentando bons resultados para o problema de agrupamento. Em parte, a explicação se deve ao fato desta técnica ser de fácil implementação e de ter complexidade de execução igual a $O(n)$, onde $n$ é o número de padrões submetidos ao algoritmo.

A idéia central do *k-means* consiste na maximização da distância inter-grupos, de mesmo modo que minimiza a distância intra-grupo. A precisão da classificação pode ser avaliada comparando-se a variabilidade intra-grupo (que é pequena se a classificação é boa) e a variabilidade inter-grupos. Os valores obtidos devem ser inversamente proporcionais, ou seja, uma classificação é considerada boa quando a variação intra-grupo for pequena e a inter-grupos grande.

Inicialmente, o algoritmo distribui aleatoriamente os elementos de um certo conjunto de dados em $k$ grupos, onde $k$ é determinado previamente, e calcula as médias dos elementos de cada grupo, que correspondem aos centróides. Em seguida, cada elemento é deslocado para o grupo correspondente ao centróide mais próximo. Com esse novo re-arranjo dos elementos, novos centróides são calculados. O processo é repetido até que nenhuma mudança ocorra, ou seja, os grupos se estabilizem (nenhum elemento é realocado para outro grupo distinto do grupo que ele pertence). Este algoritmo visa minimizar o erro quadrático médio, definido a seguir.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2,$$

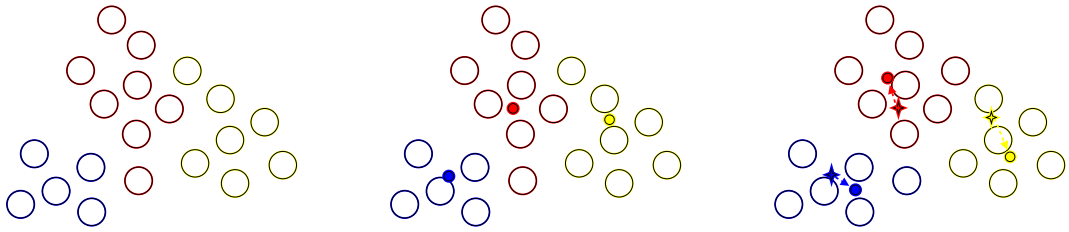
onde $||x_i^{(j)} - c_j||^2$ representa a medida de distância entre o elemento $x_i^{(j)}$ e o centróide do grupo $c_j$.

O algoritmo em alto nível é listado a seguir.

**Algoritmo *k-means***

1: Fixe o número de grupos desejado em $k$;
2: Divida o conjunto de dados aleatoriamente nos $k$ grupos;
3: Calcule o centróide de cada grupo, isto é, o ponto médio do grupo;
4: Para cada dado, calcule a distância em relação ao centróide de cada grupo;
5: Transfira o dado para o grupo cuja distância ao centróide é mínima;
6: Repita os passos (3), (4) e (5) até que nenhum dado seja mais transferido.

Na Figura 2.10, é ilustrada a aplicação dos primeiros passos do algoritmo *k-means*, considerando um conjunto com 20 dados e $k = 3$.



(a) Divisão dos dados. (b) Cálculo dos centróides. (c) Transferência dos dados.

Figura 2.10: Ilustração do algoritmo *k-means*. Em (a), os 3 grupos são gerados aleatoriamente. Em (b), os centróides dos grupos são calculados. Em (c), cada padrão é associado ao centróide do grupo mais próximo e os novos centróides são recalculados.

## 2.4 Considerações

Neste capítulo, foram abordados conceitos fundamentais que são utilizados neste trabalho. Os conceitos apresentados foram aplicados principalmente para descrição e classificação dos quadros do vídeo, essenciais ao processo de sumarização proposto. A literatura pertinente apresenta diversas técnicas de extração de características e de classificação. Entretanto, visando alcançar um dos objetivos do presente trabalho (desenvolver uma abordagem simples para a elaboração de resumos estáticos), técnicas simples foram empregadas no processo de sumarização proposto.

# Capítulo 3

# Sumarização Automática de Vídeos

Tipicamente, as soluções encontradas na literatura para o problema de sumarização são divididas em duas fases: primeiro, o vídeo é segmentado e em seguida os quadros-chave são extraídos, de acordo com algum critério. Neste capítulo, é apresentada uma revisão teórica sobre o processo de sumarização automática de vídeos, especificamente voltada para a geração de resumos estáticos. O capítulo está organizado da seguinte forma: na Seção 3.1, são apresentadas algumas formas de segmentação temporal de vídeo que são utilizadas como passo inicial para a obtenção de resumos estáticos; na Seção 3.2, são descritas as principais técnicas de extração de quadros-chave; e na Seção 3.3, são abordados os métodos para avaliação de resumos estáticos.

## 3.1 Segmentação Temporal de Vídeo

A segmentação temporal de vídeo é, geralmente, o primeiro passo para a sumarização de vídeos. Esta tem como objetivo separar o vídeo em seus componentes básicos (tomadas, quadros, por exemplo) (Koprinska e Carrato, 2001). Na literatura, a segmentação do vídeo em tomadas, mais conhecida como detecção dos limites de tomadas (do inglês, *shot boundary detection*) (Lienhart, 2001; Hanjalic, 2002), é largamente utilizada como etapa inicial na geração de resumos em diversos trabalhos, tais como (Yeung e Liu, 1995; Zhuang et al., 1998; Hanjalic e Zhang, 1999; Dufaux, 2000; Li et al., 2003; Porter et al., 2003; Rong et al., 2004; Ciocca e Schettini, 2005; Li et al., 2005; Ma e Zhang, 2005; Hadi et al., 2006; Chang e Chen, 2007; Cámara-Chávez et al., 2008). Outro tipo de segmentação é a separação do vídeo em quadros. Esse tipo de segmentação é aplicada em alguns trabalhos (Uchihashi et al., 1999; Sun e Kankanhalli, 2000; Gong e Liu, 2000; Yahiaoui et al., 2001; Mundur et al., 2006; Furini et al., 2007). Os dois tipos de segmentação são detalhados a seguir.

### 3.1.1 Detecção dos Limites de Tomadas

Para identificar os limites de uma tomada, é necessário detectar a transição entre duas tomadas consecutivas. Os tipos de transições são divididos em transições abruptas (cortes) e transições graduais (*fade-out*, *fade-in*, *dissolve* e *wipe*).

**corte** O corte (Figura 3.1) é a forma mais simples de transição entre duas tomadas, e é resultante da mudança instantânea de uma tomada para outra.

***fade-out*** Consiste na diminuição progressiva da luminosidade nos quadros de uma tomada até a obtenção de quadros monocromáticos (Figura 3.2).

***fade-in*** A transição *fade-in* (Figura 3.3) consiste no aumento progressivo da luminosidade, a partir de quadros monocromáticos, até a visualização da tomada, com sua luminosidade natural.

***dissolve*** É uma generalização da transição *fade*, onde ocorre a sobreposição dos últimos quadros da tomada anterior com os primeiros quadros da tomada seguinte, através de *fade-out* seguido por um *fade-in* (Figura 3.4).

***wipe*** Neste tipo de transição, os quadros de uma tomada são substituídos pelas regiões equivalentes da tomada seguinte (Figura 3.5).

Diversas soluções para segmentação de vídeo em tomadas têm sido propostas na literatura (Guimarães et al., 2003; Cámara-Chávez et al., 2007). Amplas revisões podem ser encontradas em (Boreczky e Rowe, 1996; Koprinska e Carrato, 2001; Lienhart, 2001; Hanjalic, 2002; Cotsaces et al., 2006; Yuan et al., 2007). Para segmentar o vídeo em tomadas, a maioria dos métodos utiliza informações visuais presentes nas cenas que compõem a tomada, como: histograma de cor, informações de bordas de objetos na imagem (Chang e Chen, 2007); coeficientes de transformadas (transformada discreta de Fourier, transformada discreta do cosseno, transformada wavelet) (Yeung e Liu, 1995; Li et al., 2003), e análise de movimentos (Dufaux, 2000; Porter et al., 2003; Ma e Zhang, 2005).

Atualmente, a detecção de transições abruptas têm sido realizada com sucesso, diferentemente da detecção de transições graduais que continua sendo um desafio, haja visto a quantidade de soluções propostas na literatura. A maioria das abordagens geralmente falham quando a ação acontece durante a transição gradual. Segundo Cotsaces et al. (2006), a melhoria das soluções para este problema irá envolver cada vez mais uma inclusão rigorosa da informação *a priori* sobre o processo físico de formação da tomada, como também uma análise teórica aprofundada sobre as transições.

### 3.1.2 Separação em Quadros

Neste tipo de segmentação, não há análise temporal do vídeo. Cada quadro é tratado separadamente, o vídeo é decomposto em um conjunto de imagens. Após realizada a decomposição, procede-se com a análise para a identificação dos quadros-chave. No contexto da sumarização automática de vídeos, Uchihashi et al. (1999); Sun e Kankanhalli (2000) consideram todos os quadros para criar os resumos, diferentemente de Gong e Liu (2000); Yahiaoui et al. (2001); Rao et al. (2004); Mundur et al. (2006); Furini et al. (2007) que utilizam uma pré-amostragem (*pre-sampling*) dos quadros, de acordo com algum parâmetro pré-especificado.

Figura 3.1: Ilustração da transição *corte* do vídeo *The Great Web of Water, segment 01* (disponível via *Open Video*).

Figura 3.2: Ilustração da transição *fade-out* do vídeo *Oceanfloor Legacy, segment 04* (disponível via *Open Video*).

Figura 3.3: Ilustração da transição *fade-in* do vídeo *America's New Frontier, segment 03* (disponível via *Open Video*).

Figura 3.4: Ilustração da transição *dissolve* do vídeo *A New Horizon, segment 01* (disponível via *Open Video*).


Figura 3.5: Ilustração da transição *wipe* do vídeo *A New Horizon, segment 06* (disponível via *Open Video*).

A técnica de amostragem reduz o número de quadros do vídeo a serem analisados. Conseqüentemente, quanto menor for a taxa de amostragem, menor será o tempo de execução para criar os resumos. No entanto, dependendo das características do vídeo, a qualidade dos resultados pode ser comprometida. Vídeos com tomadas muito longas tendem a ter vantagem com a técnica de pré-amostragem, por outro lado, vídeos com tomadas muito curtas podem vir a sofrer com a não representação de parte importante de seu conteúdo. Esta relação entre *perda de informação* e o *tamanho da tomada* está diretamente associada ao parâmetro utilizado para selecionar as amostras a serem consideradas durante a etapa de sumarização.

## 3.2 Técnicas de Extração de Quadros-chave

O passo seguinte à segmentação é a extração de quadros-chave. Segundo Truong e Venkatesh (2007), as técnicas de extração de quadros-chave podem ser classificadas em oito categorias: 1) mudança significativa do conteúdo, 2) variância temporal semelhante, 3) máxima representatividade do quadro, 4) agrupamento, 5) correlação mínima entre quadros-chave, 6) erro de reconstrução de seqüência, 7) simplificação de curvas, e 8) eventos "interessantes". Nas seções seguintes é abordada cada categoria listada.

### 3.2.1 Mudança Significativa do Conteúdo (*Sufficient Content Change*)

O método analisa seqüencialmente os quadros do vídeo, selecionando um quadro como quadro-chave se o seu conteúdo visual apresentar uma diferença significativa do quadro-chave selecionado anteriormente (veja a Figura 3.6).

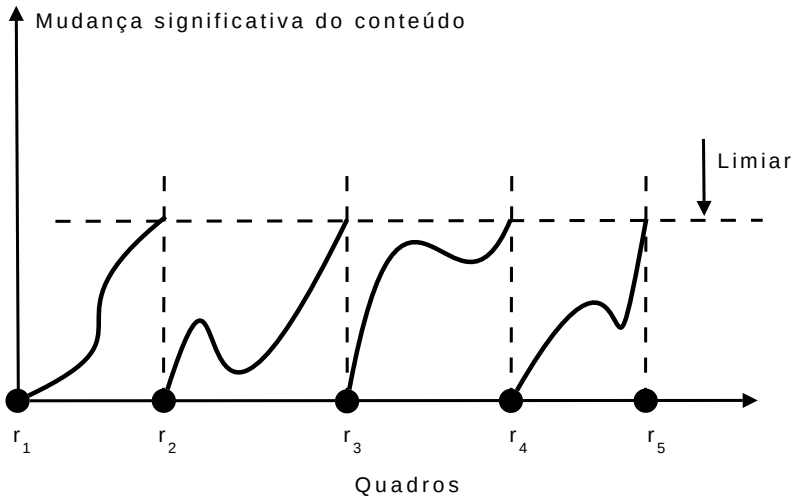



Figura 3.6: Ilustração do método mudança significativa do conteúdo. (Truong e Venkatesh, 2007).

Para medir a mudança significativa do conteúdo do vídeo, diversas métricas têm sido propostas na literatura, sendo que a mais popular é a diferença entre histogramas utilizada em (Yeung e Liu, 1995; Zhang et al., 1997; Kang et al., 1999). Algumas métricas alternativas

propostas são o cálculo de mudanças nas propriedades geométricas de objetos extraídos do vídeo (Kim e Hwang, 2002) e a função de energia acumulada, computada a partir do deslocamento de uma janela sobre dois quadros sucessivos (Zhang et al., 2003).

O método mudança significativa do conteúdo pode ser utilizado em aplicações em tempo-real e/ou *online* por ser simples e imediato. Entretanto, apesar do processamento seqüencial do vídeo garantir que a extração dos quadros-chave seja feita de maneira eficiente, o conjunto de quadros-chave gerado pode apresentar redundâncias. Um conjunto de quadros-chave mais conciso é obtido em (Rasheed e Shah, 2003), onde um quadro é selecionado como quadro-chave se o quadro for significativamente diferente de todos os quadros-chave previamente selecionados.

### 3.2.2 Variância Temporal Semelhante (*Equal Temporal Variance*)

Esta abordagem, variante do método mudança significativa do conteúdo, requer que o número de quadros-chave seja definido previamente. Além disso, o método da *variância temporal semelhante* supõe que um conjunto de quadros-chave aceitável deve ter quadros que individualmente representam segmentos do vídeo com variância temporal semelhante.

Em (Lee e Kim, 2002; Divakaran et al., 2002; Fauvet et al., 2004), a função de variância temporal é obtida por meio da aproximação do conteúdo cumulativo que varia de acordo com a seqüencia dos quadros do vídeo. Enquanto que em (Sun e Kankanhalli, 2000), a função de variância temporal é expressa por meio da diferença entre o primeiro e o último quadro do segmento.

Uma vez que os limites dos segmentos tenham sido determinados, Divakaran et al. (2002); Fauvet et al. (2004) simplesmente selecionam o primeiro quadro e o quadro central do segmento para representar o quadro-chave, respectivamente. Em (Lee e Kim, 2002), um quadro-chave é selecionado de cada segmento por meio da minimização da diferença total entre o quadro-chave e todos os quadros dentro do mesmo escopo.

Em geral, o método variância temporal semelhante apresenta um custo computacional maior que o método mudança significativa do conteúdo.

### 3.2.3 Máxima Representatividade do Quadro (*Maximum Frame Coverage*)

Proposto por Chang et al. (1999), e também conhecido como uma abordagem baseada em um limite de confiança, este método tenta obter a representatividade dos quadros (ou seja, a lista de quadros que um dado quadro pode representar) para gerar o conjunto de quadros-chave.

Em (Chang et al., 1999), a representatividade do quadro-chave é composta por todos os quadros do vídeo que apresentam semelhança visual em relação ao quadro-chave. Os autores visam encontrar uma solução aproximada através de um algoritmo guloso, o qual extrai os quadros de máxima representatividade a cada passo. Esta iteração ocorre até que todos os quadros tenham sido analisados.

O princípio da máxima representatividade do quadro também é a base da solução proposta por Yahiaoui et al. (2001). Entretanto, a representatividade de um quadro é o número de

fragmentos de tamanho fixo (e não de quadros individuais) que contém pelo menos um quadro similar a este quadro. Uma técnica de programação dinâmica é utilizada para selecionar um número de quadros pré-especificados como quadros-chave. No entanto, a vantagem de utilizar fragmentos de tamanho fixo em vez de uma lista de quadros não é clara e não é justificada neste trabalho. Rong et al. (2004); Cooper e Foote (2005) fazem uma analogia entre a extração de quadros-chave e a extração de palavras-chave para recuperação de informação textual através do método TF–IDF (do inglês, *Term Frequency – Inverse Document Frequency*). Assim, para um quadro ser considerado como um quadro-chave de uma tomada/cena, é necessário que este tenha uma boa representatividade da tomada/cena, mas uma baixa representatividade do vídeo.

Este método apresenta como vantagem o fato de não exigir que o quadro represente um segmento contínuo do vídeo como nos métodos mudança significativa do conteúdo e variância temporal semelhante, e portanto, pode gerar um conjunto de quadros-chave mais conciso. No entanto, a dissimilaridade precisa ser computada para todo par de quadros, o que o torna mais caro computacionalmente e, por isso, inadequado para aplicações em tempo-real e/ou *online*. Além disso, o método exige uma métrica precisa de similaridade visual para fornecer uma solução ótima.

### 3.2.4 Agrupamento (*Clustering*)

A técnica de agrupamento consiste na classificação não-supervisionada de padrões em grupos (*clusters*) (Jain et al., 1999). Esta abordagem trata os quadros como pontos no espaço de características e supõe que os pontos representativos dos grupos gerados neste espaço podem representar os quadros-chave do vídeo. A abordagem geralmente é divida em quatro passos, listados a seguir. Os métodos existentes podem ser diferenciados pela forma como estes passos são implementados.

**Pré-processamento de dados** O pré-processamento dos dados de entrada tem como objetivo melhorar a eficácia e a eficiência do processo de agrupamento. Xiong et al. (1997) extraem os possíveis quadros-chave através do método mudança significativa do conteúdo e os utiliza como entrada do algoritmo de agrupamento. Do mesmo modo, Girgensohn e Boreczky (1999) usam como entrada um número pré-definido de quadros, os quais são largamente diferentes de pelo menos um dos quadros adjacentes. Em (Gibson et al., 2002) e (Yu et al., 2004), cada quadro é transformado em auto-espaços através da análise de componentes principais (do inglês, *Principal Component Analysis* – PCA) e do *kernel* PCA, respectivamente. O objetivo desta transformação é reduzir a dimensão do espaço dos dados, mantendo as variações significativas dos dados originais.

**Agrupamento de dados** As técnicas comuns de agrupamento e suas variações podem ser aplicadas aos dados para identificar possíveis grupos. Xiong et al. (1997); Zhuang et al. (1998) usam uma técnica de agrupamento seqüencial que atribui o quadro a um grupo existente se a similaridade entre eles for máxima e superior a um limiar, e cria um novo grupo caso contrário. Girgensohn e Boreczky (1999); Ciocca e Schettini (2006) aplicam

o algoritmo hierárquico *complete-link*, Hanjalic e Zhang (1999) utilizam o algoritmo particional *k-means*, Yu et al. (2004) empregam o algoritmo de agrupamento *fuzzy c-means*, Hadi et al. (2006) usam o algoritmo *k-medoid*. Gibson et al. (2002) utilizam modelos de mistura gaussiana (do inglês, *Gaussian Mixture Models* – GMM), onde o número de componentes (ou seja, o número de quadros-chave) que melhor representa os grupos é calculado através de um processo iterativo baseado em um limiar de semelhança máxima. Mundur et al. (2006) geram os grupos através da triangulação de Delaunay. Furini et al. (2007) utilizam uma versão melhorada do algoritmo *furthest-point-first* para obter os grupos.

**Filtragem de grupos** Uma vez que os grupos resultantes podem apresentar ruídos e/ou o próprio grupo pode não ser suficientemente significativo, os quadros-chave não são extraídos de todos os grupos. Zhuang et al. (1998) consideram apenas os grupos que sejam maiores que a média do tamanho dos grupos. Girgensohn e Boreczky (1999) sugerem que o grupo deve conter pelo menos uma seqüência ininterrupta de quadros com uma duração mínima para garantir que artefatos do vídeo e outros eventos não-significativos não sejam representados pelos quadros-chave. Em (Uchihashi et al., 1999), a filtragem dos grupos é realizada através de uma pontuação computada a partir do tamanho dos grupos e suas características comuns.

**Extração dos pontos representativos de cada grupo** A solução mais comum e intuitiva é seleção do ponto mais próximo do centróide do grupo como o ponto representativo do grupo, gerando assim o conjunto de quadros-chave do vídeo (Zhuang et al., 1998; Hanjalic e Zhang, 1999; Uchihashi et al., 1999; Gong e Liu, 2000; Yu et al., 2004; Hadi et al., 2006; Mundur et al., 2006; Furini et al., 2007). Similarmente, os centróides dos grupos são os centros de cada componente GMM (Gibson et al., 2002). Em (Girgensohn e Boreczky, 1999), os quadros mais representativos de cada grupo são selecionados através de restrições temporais.

De acordo com Truong e Venkatesh (2007), os métodos de agrupamento são bastante empregados na extração dos quadros-chave devido ao seu uso comum em análise de dados. Entretanto, uma vez que o significado semântico dos grupos dos dados do vídeo geralmente apresentam alta variância intra-classe e baixa variância inter-classe, a obtenção de sucesso na extração é muito difícil. Trabalhos nessa área de pesquisa muitas vezes não conseguem avaliar adequadamente os resultados do processo de agrupamento. Além disso, as técnicas de agrupamento geralmente não são apropriadas quando é necessário preservar a ordem temporal do vídeo.

### 3.2.5 Correlação Mínima entre Quadros-chave (*Minimum Correlation Among Keyframes*)

Técnicas nessa abordagem geralmente lidam com a restrição da taxa de extração de quadros-chave e trabalham com a idéia que o conjunto de quadros-chave deve ter uma correlação

mínima entre seus elementos (algumas vezes é considerada somente a correlação entre elementos sucessivos). O objetivo é selecionar os quadros que não são similares entre si. Diferentes algoritmos são utilizados para obter uma solução quase ótima, tais como pesquisa logarítmica (Doulamis et al., 1998), busca estocástica (Avrithis et al., 1999), algoritmo genético (Doulamis et al., 2000) e algoritmo guloso (Liu e Kender, 2002b).

Apesar do critério da correlação mínima garantir um baixo nível de redundância dos quadros-chave, este método não apresenta bons resultados na presença de *outliers*. Porter et al. (2003) tentou resolver este problema representando o conjunto de quadros-chave como um caminho em um grafo ponderado direcionado, no qual cada quadro do vídeo corresponde a um vértice do grafo e as arestas representam a correlação entre dois quadros, que é computada a partir de vetores dos movimentos. Este método requer que o primeiro e o último quadro pertençam ao conjunto de quadros-chave e a solução ótima, portanto, equivale ao caminho mais curto entre o primeiro e o último vértice/quadro.

### 3.2.6 Erro de Reconstrução de Seqüência (*Sequence Reconstruction Error – SRE*)

Esta técnica, também conhecida como *erro de reconstrução da tomada*, mede a capacidade do conjunto de quadros-chave reconstruir a seqüência/tomada original do vídeo. Geralmente, o método é aplicado quando o número de quadros-chave é definido previamente e a progressão temporal é importante para uma aplicação específica.

Seja $\mathcal{I}(t, \mathcal{R})$ uma função de interpolação de quadros aplicada para a imagem inteira ou algumas características da imagem em um determinado tempo $t$ na seqüência do vídeo $V$ do conjunto de quadros-chave $\mathcal{R}$. O SRE é definido como:

$$\mathcal{E}(V, \mathcal{R}) = \sum_{i=1}^{n} \mathcal{D}(f_i, \mathcal{I}(i, \mathcal{R})), \tag{3.1}$$

onde $f_i$ é o $i$-ésimo quadro do vídeo e $\mathcal{D}$ é a função que calcula a diferença entre dois quadros.

Quanto melhor for a reconstrução da seqüência/tomada do vídeo obtida através da interpolação dos quadros-chave, mais próximo da seqüência/tomada original do vídeo será o resultado. Da mesma forma, quanto melhor o algoritmo de seleção do conjunto de quadros-chave, maior será a capacidade de sumarização deste conjunto.

Entretanto, para encontrar uma solução prática, certas suposições são feitas para a função de interpolação de quadros $\mathcal{I}$ e a medida de diferença entre quadros $\mathcal{D}$. Em (Hanjalic et al., 1998; Liu e Kender, 2002a; Lee e Kim, 2003), cada quadro-chave $f_{r_i}$ representa o segmento $(b_i, b_{i+1} - 1)$ do vídeo, e a função de interpolação é definida como:

$$\mathcal{I}(i, \mathcal{R}) = f_{r_i} \Longleftrightarrow b_t \leq i \leq b_{t+1}. \tag{3.2}$$

A função de erro de reconstrução da seqüência $\mathcal{E}(V, \mathcal{R})$, ilustrada pela região sombreada

na Figura 3.7, é definida como:

$$\mathcal{E}(V,\mathcal{R}) = \sum_{i=1}^{n} \mathcal{D}(f_i, \mathcal{I}(i,\mathcal{R})) = \sum_{i=1}^{k} \sum_{j=b_i}^{b_{i+1}-1} \mathcal{D}(f_j, f_{r_i}), \tag{3.3}$$

onde $k$ representa o número de quadros-chave (definido previamente) e a diferença absoluta de uma determinada característica é utilizada para medir a diferença entre duas imagens.

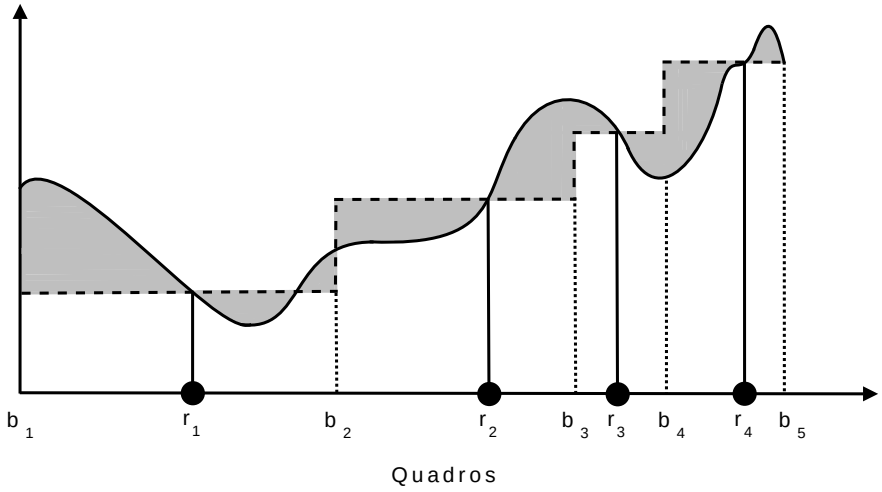



Figura 3.7: Ilustração do método Erro de Reconstrução da Seqüência (Truong e Venkatesh, 2007).

Segundo Li et al. (2004), uma solução ótima pode ser encontrada através de programação dinâmica se for suposto que cada quadro-chave representa apenas os quadros seguintes. Então, supondo que a solução ótima é obtida para um conjunto de quadros-chave de qualquer tamanho, um algoritmo de busca por bissecção pode ser aplicado para encontrar o conjunto ótimo de quadros-chave, dado um determinado limite.

Uma solução alternativa é proposta por Liu et al. (2004), onde em vez de usar um quadro-chave para interpolar com os quadros, são utilizados dois quadros-chave adjacentes.

### 3.2.7 Simplificação de Curvas (*Curve Simplification*)

A abordagem *simplificação de curvas* está relacionada com as técnicas de agrupamento, uma vez que também trata cada quadro do vídeo como um ponto em um espaço multidimensional. No entanto, em vez de formar grupos, esta abordagem procura por um conjunto de pontos tal que a remoção de pelo menos um ponto altera a forma da curva que liga todos os pontos através da ordem temporal destes. Alguns algoritmos padrão de simplificação de curvas são aplicados por DeMenthon et al. (1998); Latecki et al. (2001); Ćalić e Izquierdo (2002).

Na Figura 3.8, é ilustrada essa abordagem. Seis quadros-chave ótimos são selecionados a partir do método de simplificação de curvas. Se o quadro-chave $q_2$ for substituído pelo

quadro $q_3$, este não representará uma aproximação da curvatura tão bem quanto quadro-chave original, e por conseguinte, toda a curva será menos representativa.

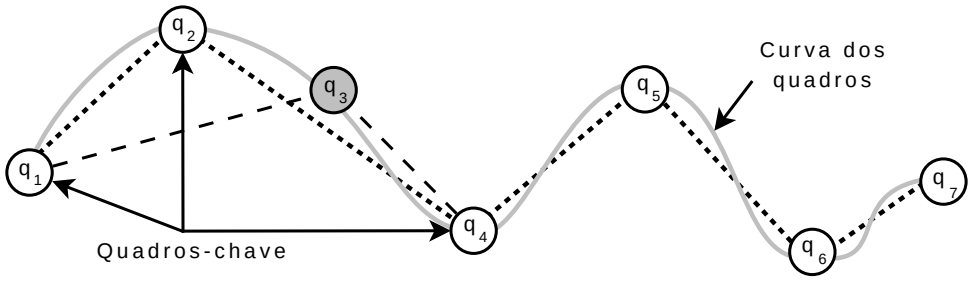



Figura 3.8: Ilustração do método simplificação de curvas (Truong e Venkatesh, 2007).

Assim como as abordagens mudança significativa do conteúdo, variância temporal semelhante, e erro de reconstrução da tomada, os quadros-chave obtidos através da abordagem simplificação de curvas dividem o vídeo em segmentos de quadros contínuos, preservando a ordem temporal. Entretanto, além da abordagem não ser eficiente, o conjunto gerado pode apresentar redundâncias.

### 3.2.8 Eventos "Interessantes" (*"Interesting" Events*)

Métodos baseados em eventos "interessantes" tentam identificar os quadros que contêm conteúdo semântico relevante. A maioria dos trabalhos nesta categoria supõe uma associação entre o "grau de importância" do quadro e os padrões de movimento nas proximidades do quadro, e também conteúdos específicos no vídeo (por exemplo, faces humanas).

Dufaux (2000) propõem uma abordagem para o problema específico de extrair um único quadro-chave de todo o vídeo. Inicialmente, a seleção da tomada é efetuada de acordo com alguns critérios, tais como a duração da tomada, a atividade espacial e a probabilidade da presença de pessoas. Uma vez que a tomada mais representativa seja selecionada, Dufaux (2000) sugere que o quadro-chave da tomada mais representativa seja selecionado de acordo com a baixa movimentação (para evitar artefatos gerados da codificação do vídeo), alta atividade espacial e alta probabilidade da presença de pessoas.

A extração dos quadros-chave em (Liu e Kender, 2002c) é restrita aos vídeos educacionais e concentra-se em um tipo de cena específica: notas de aulas manuscritas. Particularmente, o conteúdo do vídeo, e conseqüentemente a importância de um quadro, é medido através da quantidade de "pixels de tinta" do quadro. Um quadro é selecionado como quadro-chave se o quadro for claro, isto é, apresenta pequena variação do conteúdo dentro de uma janela de tempo, e se for suficientemente diferente de todos os outros quadros claros.

Em (Liu et al., 2003), a extração dos quadros-chave é baseada nos padrões de movimento de uma tomada. Os autores desenvolvem um modelo triangular de detecção da energia do movimento, no qual uma tomada é dividida em subsegmentos de consecutivos padrões de movimento de acordo com o processo de aceleração e desaceleração do movimento. Neste

modelo, os quadros nos vértices do triângulos são selecionados como quadros-chave. Do mesmo modo, Han e Kweon (2005) extraem os quadros-chave nos pontos de curvatura do gráfico descrito pelo movimento da câmera. Entretanto, este modelo analisa a magnitude e a direção do movimento.

Embora estas técnicas possam funcionar bem para algumas configurações experimentais específicas, o problema com estes métodos é que eles dependem fortemente de heurísticas extraídas de observações empíricas de uma base de dados restrita. A complexidade e a irregularidade nos padrões de movimento de uma tomada também podem tornar estas soluções ineficazes em seqüências não-testadas ou em domínios específicos.

## 3.3 Métodos para Avaliação de Resumos

Para obter avanço na área de sumarização automática de vídeos, ou em qualquer outro campo de pesquisa, a eficácia e/ou eficiência de uma nova solução deve ser avaliada, principalmente em relação aos métodos existentes. No entanto, a área de sumarização de vídeo não dispõe de um modelo de avaliação consolidado e a maioria das formas de avaliação propostas apresentam graves deficiências, resultando em trabalhos com os seus próprios métodos de avaliação, sendo que na maioria das vezes não são realizadas comparações com as técnicas existentes. Em parte, isto ocorre porque, diferentemente das outras áreas de pesquisa, tais como detecção e reconhecimento de objetos, a avaliação de resumos de vídeos não é uma tarefa direta devido à inexistência de um *ground truth*.

Segundo Truong e Venkatesh (2007), as principais formas de avaliação de resumos são agrupadas em três categorias: descrição dos resultados, métricas objetivas e avaliação através de usuários.

### 3.3.1 Descrição dos Resultados

A descrição dos resultados é a forma mais simples de avaliação, pois não realiza comparações com outras técnicas. Neste tipo de avaliação, os resumos gerados, que podem ser descritos e/ou apresentados visualmente, são utilizados para indicar se o método é adequado para a sumarização de vídeos. Além disso, o método de descrição dos resultados é utilizado também para discutir a influência dos parâmetros do sistema ou a dinâmica do vídeo no conjunto de quadros-chave extraído (Zhuang et al., 1998; Hanjalic et al., 1998; Zhang et al., 2003; Yu et al., 2004). Alguns trabalhos tentam, de forma descritiva, explicar e ilustrar as vantagens de seus respectivos métodos comparando-os com outros existentes (Joshi et al., 1998; Vermaak et al., 2002). Esse método de avaliação demonstra ser bastante subjetivo e dificilmente fornece justificativa sólida da validade da técnica de sumarização proposta.

### 3.3.2 Métricas Objetivas

Na avaliação por métricas, para técnicas de extração de quadros-chave, a métrica é, geralmente, uma função de confiança computada a partir do conjunto de quadros-chave extraídos

e da seqüência original dos quadros (Chang et al., 1999; Liu e Kender, 2002b). Essa métrica é utilizada para comparar o conjunto de quadros-chave produzido por diferentes métodos, ou por um único método, mas com diferentes parâmetros.

Comumente, a qualidade do resumo é também medida em termos de precisão (*precision*) e revocação (*recall*). Em (Lee e Hayes, 2003), a precisão ($P$) e a revocação ($R$) são definidas, respectivamente, como:

$$P = \frac{n_C}{n_C + n_I} \quad \text{e} \quad R = \frac{n_C}{n_C + n_N},$$

onde $n_C$ representa o número de quadros-chave corretamente selecionados, $n_I$ o número de quadros-chave incorretos (falsos negativos) e $n_N$ o número de quadros-chave não selecionados do conjunto de quadros-chave correto (falsos positivos). Esse conjunto é obtido a partir dos quadros-chave gerados automaticamente e da eliminação manual dos quadros-chave similares. O número de quadros-chave resultante corresponde ao conjunto de quadros-chave corretos.

Além disso, alguns trabalhos definem as métricas de avaliação, como ocorre em (Gong e Liu, 2003; Fauvet et al., 2004; Mundur et al., 2006). Por exemplo, Mundur et al. (2006) definem três métricas para avaliar a qualidade dos resumos: fator de importância, fator de compressão e fator de sobreposição. A primeira métrica denota a importância do conteúdo representado por cada grupo através do número de quadros do grupo. A segunda métrica quantifica a relação entre o número de quadros do resumo e do vídeo. A terceira métrica é utilizada para fazer uma comparação significativa entre os resumos gerados pela técnica proposta e os resumos do Open Video.

A avaliação por métricas objetivas é bastante aplicada para medir a qualidade dos resumos. Contudo, não há uma justificativa experimental de que estas métricas sejam tão boas quanto o julgamento humano.

### 3.3.3 Avaliação através de Usuários

A avaliação através de usuários, como o próprio nome sugere, é realizada por usuários que julgam a qualidade dos resumos gerados. Este tipo de avaliação é empregado em diversos trabalhos, tais como (Dufaux, 2000; Yahiaoui et al., 2001; Ma et al., 2002; Liu et al., 2003; Furini et al., 2007; Guironnet et al., 2007; Wang et al., 2007).

Em (Dufaux, 2000) o quadro-chave, que representa todo o vídeo, é classificado como "bom", "satisfatório" ou "ruim". Este método demonstrou ser mais eficaz do que a seleção do quadro central do vídeo como quadro-chave. De modo semelhante, em (Ma et al., 2002; Liu et al., 2003), os quadros-chave são avaliados como ("bom", "neutro" ou "ruim") e ("bom", "aceitável" ou "ruim"), respectivamente. Em (Ma et al., 2002), 20 usuários avaliam cada quadro-chave de cada tomada. O método proposto não foi comparado com outras técnicas. Diferentemente, em (Liu et al., 2003), 10 usuários avaliam o conjunto de quadros-chave extraído de cada tomada. Estes são questionados se o conjunto representam bem o conteúdo da tomada. A técnica proposta também não foi comparada com outros métodos.

No método de avaliação proposto por Yahiaoui et al. (2001), inicialmente o resumo de um vídeo é apresentado aos usuários. Em seguida, um segmento do vídeo utilizado, escolhido aleatoriamente, é mostrado aos usuários. Estes são questionados se o segmento originou o resumo. A resposta deve ser positiva se pelo menos uma imagem do segmento é similar a uma imagem do resumo. A probabilidade de uma resposta correta é usada para medir a qualidade do resumo. O método de avaliação não foi comparado com outros métodos.

Furini et al. (2007) avaliam seus resumos através de um conjunto de 20 pessoas que devem responder se o resumo gerado representa bem o vídeo original. Para isso, é proposta uma escala de 1 a 5 (1 = "ruim", 2 = "pobre", 3 = "satisfatório", 4 = "bom", 5 = "excelente"). A média da opinião das pessoas indica a qualidade do resumo. Os autores comparam os seus resultados com os resumos apresentados em (Mundur et al., 2006) e os resumos disponíveis do Open Video.

Wang et al. (2007) apresentam uma avaliação similar ao método proposto por Furini et al. (2007). Os autores também utilizam uma escala de 1 (definitivamente "NÃO") a 5 (definitivamente "SIM") para medir a qualidade do resumo. Neste caso, os resumos são avaliados através de um conjunto de 10 estudantes, os quais respondem a diversas questões. Para cada vídeo são gerados quatro tipos de resumos: *Pictorial* (Yeung e Leo, 1997), *Booklet* (Hua et al., 2005), *Grid View* e *Video Collage* (técnica proposta).

As formas propostas de avaliação através de usuários citadas anteriormente apresentam intrinsecamente um alto grau de subjetividade, pois consistem no fato de estimar a qualidade do resumo de acordo com uma escala definida. Ou seja, os usuários são questionados se o resumo "está bom" ou "não está bom". Com a finalidade de reduzir essa subjetividade, Guironnet et al. (2007) propõem que os usuários criem o resumo do vídeo manualmente, em vez de avaliar os resumos gerados por alguma técnica. A partir dos resumos montados pelos usuários (neste caso, 12 usuários), um resumo "ótimo" é construído automaticamente e este é comparado com os resumos gerados por diferentes técnicas.

Dentre as formas de avaliação, a avaliação através de usuários é a mais realista, principalmente quando os quadros-chave são extraídos para tarefas que implicam interação com usuário, tal como navegação. No entanto, ainda não é a mais amplamente aplicada devido à dificuldade de estruturá-la e de tratar a subjetividade inerente a este tipo de avaliação.

## 3.4 Considerações

O processo de produção automática de resumos estáticos de vídeos é composto, basicamente, por duas etapas: 1) segmentação do vídeo e 2) extração dos quadros-chave. Durante as duas últimas décadas, diferentes técnicas foram desenvolvidas para solucionar o problema da sumarização automática de vídeos. Neste capítulo foram revistas as principais técnicas e foram listadas suas vantagens e desvantagens. Além disso, também foram abordados os métodos para avaliação de resumos encontrados na literatura. A metodologia proposta, apresentada no próximo capítulo, integra as vantagens dos conceitos empregados por alguns trabalhos e propõem soluções para as deficiências observadas.

# Capítulo 4

# Metodologia Proposta

Neste capítulo, é descrita a metodologia aplicada no presente trabalho para gerar resumos estáticos de vídeos e os métodos de avaliação propostos para estimar a qualidade dos resumos produzidos. O capítulo está organizado da seguinte forma: na Seção 4.1, é descrita a abordagem proposta para sumarização de vídeos e cada um dos seus passos é detalhado nas subseções seguintes; e na Seção 4.2, são apresentadas as formas de avaliação de resumos propostas.

## 4.1 Método Proposto para Sumarização de Vídeo

Na Figura 4.1, é apresentado o esquema geral do método proposto para sumarização de vídeo. Inicialmente, o vídeo é segmentado em quadros (passo 1). No passo 2, são extraídas as características visuais presentes nos quadros, que serão utilizadas para a descrição do conteúdo do vídeo. Nessa etapa, é utilizada uma subamostra dos quadros que compõem o vídeo. Além disso, os quadros inexpressivos[1] contidos na subamostra são eliminados. Em seguida, os quadros são agrupados por meio de uma abordagem de classificação não-supervisionada (passo 3). No passo 4, o quadro-chave de cada grupo é identificado. Para gerar o resumo estático (passo 5), os quadros-chave semelhantes são eliminados com o propósito de manter a qualidade do resumo. E por fim, os quadros-chave são dispostos em ordem cronológica para facilitar o entendimento visual do resultado. Cada um destes passos é detalhado a seguir.

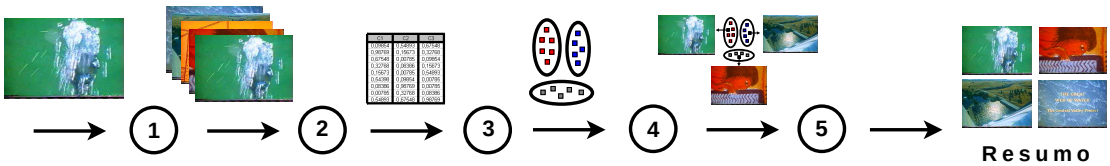



Figura 4.1: Diagrama da metodologia proposta para gerar resumos estáticos de vídeos.

[1]Quadros inexpressivos são quadros monocromáticos, que geralmente ocorrem devido a efeitos de *fade-in* ou *fade-out* ou devido ao uso de *flashes*.

### 4.1.1 Segmentação do Vídeo

Na literatura, a segmentação do vídeo em tomadas é largamente utilizada como etapa inicial na geração de resumos (veja a Seção 3.1). Entretanto, a abordagem baseada em tomadas pode apresentar redundâncias, uma vez que diferentes tomadas podem conter conteúdo similar. Por exemplo, em vídeos de notícias ou documentários, o apresentador pode aparacer várias vezes em diferentes tomadas e conseqüentemente quadros semelhantes podem compor o resumo.

Na Figura 4.2, são ilustradas diferentes tomadas para um mesmo vídeo. Pode-se notar que estas são bastante similares e que o resumo apresentará redundâncias, caso este seja formado pelos quadros-chave extraídos de cada tomada.

(a) Quadros 486 a 492

(b) Quadros 772 a 778

(c) Quadros 1090 a 1096

(d) Quadros 2522 a 2528

Figura 4.2: Diferentes tomadas do vídeo *Senses And Sensitivity, Introduction to Lecture 2* (disponível via *Open Video*).

Outro aspecto em relação à segmentação do vídeo em tomadas é a própria identificação das tomadas. Apesar do campo ter avançado bastante, a delimitação das tomadas em vídeo ainda é um problema não resolvido. Sendo assim, caso as tomadas não sejam identificadas de forma correta, a qualidade do resumo pode ser prejudicada, pois, a depender da segmentação, o resumo pode apresentar informações redundantes (por exemplo, quando uma tomada é separada em duas ou mais tomadas) ou não apresentar uma informação relevante (por exemplo, quando duas ou mais tomadas são identificadas como uma única tomada).

Diante desses fatores, em vez de segmentar os vídeos em tomadas, a abordagem empregada neste trabalho separa os vídeos em quadros. Além disso, o método proposto não analisa todos os quadros do vídeo, mas apenas um subconjunto (uma amostragem).

Como se sabe, um vídeo é composto por uma seqüência de quadros. Para assegurar que os

seres humanos não percebam qualquer descontinuidade no fluxo de quadros, é necessária uma taxa de pelo menos 25 quadros por segundo (Hammoud, 2006), o equivalente à 90.000 imagens em uma hora de vídeo. Assim, é intuitivamente óbvio que, para uma taxa de quadros de 25 fps, os 25 quadros mostrados a cada segundo contêm informação redundante. Conseqüentemente, não é necessário analisar todos os quadros do vídeo para gerar os resumos com qualidade. Diferentes taxas de amostragem são experimentadas neste trabalho. Os resultados são apresentados no Capítulo 5.

### 4.1.2 Extração de Características

Imagens podem ser descritas por características como cor, forma e textura. A cor, por ser uma característica intrínseca das imagens, possui um papel muito importante na representação destas (Stehling et al., 2002). Neste trabalho são utilizadas as características de cor da imagem. Para gerar o vetor de características são utilizadas ferramentas simples para descrever o conteúdo visual dos quadros: histograma de cor normalizado ou perfil de linha (vertical, horizontal e diagonal) (veja a Seção 2.2.1).

O histograma de cor representa a distribuição da freqüência de ocorrência dos valores cromáticos em uma imagem. Ele pode ser descrito computacionalmente através de uma estrutura de dados unidimensional. O tamanho desta estrutura deve ser igual ao número de cores possíveis no modelo de cor utilizado para representar a imagem. Considerando que imagens coloridas tipicamente têm as cores representadas em um modelo de cor com 256 níveis para cada canal de cor, o que resulta em $256 \times 256 \times 256 = 16.777.216$ cores distintas, uma redução se faz necessária para viabilizar a aplicação do algoritmo de agrupamento (próximo passo da metodologia proposta). Para isto, optou-se por fazer a quantização das cores utilizadas na representação das imagens. A quantidade de cores quantizadas para representar cada canal de cor é determinada pelo parâmetro *número de cores quantizadas*. Os testes relativos aos modelos de cores (veja a Seção 2.2.1) são realizados com os modelos de cores RGB e HSV.

O perfil de linha representa os valores cromáticos presentes em uma determinada linha da imagem. Como uma linha da imagem não é suficiente para identificar a similaridade entre as imagens, é analisado mais de um perfil de linha para cada imagem. O número de perfis de linha analisado é determinado pelo parâmetro *intervalo entre perfis de linha*. Por exemplo, se o valor do parâmetro for igual a 10, então será gerado o perfil de linha da imagem a cada 10 linhas. Os testes são realizados com o modelo de cor RGB.

### 4.1.3 Eliminação de Quadros Inexpressivos

Os quadros do vídeo que são considerados inexpressivos são os quadros monocromáticos, que geralmente ocorrem no vídeo devido a efeitos de *fades* ou devido ao uso de *flashes*.

A eliminação destes quadros é feita através de um procedimento bastante simples: o cálculo do desvio padrão dos valores do vetor de característica. Como o desvio padrão dos valores, no caso dos quadros monocromáticos, é igual a zero ou um número próximo de zero, pois pode ocorrer casos em que o quadro não seja completamente de uma mesma cor, mas pode ser

considerado como um quadro inexpressivo, basta eliminar os quadros que apresentam estes valores.

Esse passo é aplicado também no trabalho desenvolvido por Furini et al. (2007). Diferentemente do presente trabalho, que emprega esse passo como uma etapa de pré-processamento dos quadros, Furini et al. utilizam-no como pós-processamento dos resumos, eliminando os quadros inexpressivos que podem ter sido selecionados pelo algoritmo de agrupamento. No entanto, não é razoável utilizar estes quadros no processo de agrupamento, já que eles são considerados inexpressivos. Por este motivo, a eliminação destes quadros é realizada anteriormente à etapa de agrupamento neste trabalho.

### 4.1.4 Técnica de Agrupamento

Neste trabalho, foi utilizado o algoritmo de agrupamento *k-means* (MacQueen, 1967). Conforme descrito na Seção 2.3.2, este método requer a definição prévia do número de grupos ($k$). Para se determinar $k$ automaticamente, as características de cor são utilizadas. Além do procedimento proposto ser simples, ele apresenta de forma rápida uma estimativa razoável do número de grupos que melhor representa o conteúdo do vídeo.

O método proposto para determinar o número de grupos, o algoritmo *k-means* e as modificações realizadas neste são apresentados a seguir.

#### Determinação do Número de Grupos ($k$)

O número de grupos $k$ é obtido por meio da distância Euclidiana par-a-par entre os vetores de características (extraídos anteriormente) dos quadros consecutivos do vídeo. É importante ressaltar que a distância é aplicada na subamostragem de quadros (obtidos no passo inicial), em vez de todos os quadros do vídeo. A determinação de $k$ baseia-se no limiar $\tau$, que define a ocorrência de variação do conteúdo da seqüência do vídeo. Caso a distância entre quaisquer dois quadros consecutivos seja maior que $\tau$, então $k$ é incrementado. O valor de $k$ utilizado, obtido por meio de testes experimentais, foi 0,5.

Na Figura 4.3, é mostrado um exemplo de como as distâncias entre os quadros são distribuídas ao longo do tempo. Pode-se observar que existem instantes de tempo nos quais a distância entre os quadros consecutivos apresentam uma variação alta (correspondendo aos picos), enquanto existem longos períodos nos quais a variação entre as distâncias é pequena (correspondendo a regiões densas). Geralmente, os picos correspondem aos movimentos bruscos no vídeo ou a uma mudança de cena, enquanto que as regiões densas de quadros correspondem aos quadros similares. Assim, os quadros entre os picos podem ser considerados conjuntos de quadros similares e o número de picos equivale a quantidade de conjuntos. Ou seja, o número de grupos ($k$) é dado pelo número de picos.

É importante ressaltar que o método para determinação do número de grupos, apresentado neste trabalho, é baseado na forma mais comum de detecção dos limites de tomadas (Guimarães, 2003), pois o valor de $k$ é incrementando a cada variação significativa do conteúdo da seqüência do vídeo.

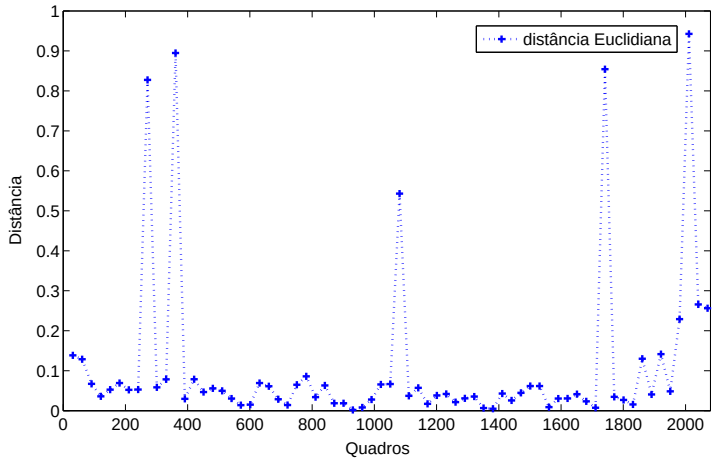



Figura 4.3: Distância entre pares consecutivos de uma amostra de quadros para o vídeo *The Great Web of Water, segment 02* (disponível via *Open Video*).

#### Algoritmo *k-means*

O *k-means* (veja Seção 2.3.2) foi adotado por ser uma abordagem simples, bastante difundida e por fornecer bons resultados no processo de classificação não-supervisionada de dados (Duda et al., 2001). No entanto, outra forma de alocação dos quadros nos $k$ grupos foi aplicada a fim de melhorar o desempenho e obter resultados mais eficazes.

Para facilitar o processo de agrupamento, os quadros são alocados nos grupos de forma seqüencial, em vez de alocá-los de forma aleatória. Por exemplo, suponha $k = 6$ e que o conjunto de amostragem do vídeo contém 60 quadros. Para distribuir os 60 quadros nos 6 grupos, os quadros são alocados igualmente de forma seqüencial por grupo. Ou seja, os primeiros 10 quadros são alocados no primeiro grupo, em seguida os próximos 10 quadros são alocados no segundo grupo e assim sucessivamente. Este procedimento é empregado porque os quadros consecutivos da seqüência muito provavelmente apresentam alguma similaridade entre si, sendo portanto bons candidatos a permanecerem no mesmo grupo.

### 4.1.5 Identificação dos Quadros-chave

Uma vez que os grupos tenham sido formados, o próximo passo consiste em identificar os quadros-chave. Antes desta identificação, uma estratégia proposta por Zhuang et al. (1998), e aplicada por Cámara-Chávez (2007), é apenas a seleção dos grupos grandes o suficiente para ser um grupo-chave. Segundo Zhuang et al. (1998), um grupo é grande o suficiente se o seu tamanho é maior que $N/k$, onde $N$ representa o número de quadros analisados do vídeo e $k$ é o número de grupos, ou seja, $N/k$ representa a média do tamanho dos grupos. Para cada grupo-chave, o quadro mais próximo do centróide do grupo-chave é selecionado como quadro-chave.

Neste trabalho, a identificação dos quadros-chave baseia-se nesta idéia. Um grupo é consi-

derado como grupo-chave se o seu tamanho é maior que $N/2k$, a metade da média do tamanho dos grupos. Este valor foi aplicado por ser um ponto de corte menos brusco que a média do tamanho dos grupos. Da mesma forma, para cada grupo-chave, o quadro mais próximo do centróide do grupo-chave é selecionado como o quadro mais representativo. Para medir a proximidade entre os quadros e o centróide é utilizada a distância Euclidiana.

Os resumos produzidos no presente trabalho são obtidos com e sem a aplicação desta estratégia de seleção dos grupos-chave. A análise dos resultados é apresentada no Capítulo 5.

### 4.1.6 Eliminação dos Quadros-chave Semelhantes

O objetivo desta etapa consiste em eliminar possíveis quadros-chave semelhantes. Para isso, os quadros-chave são comparados entre si de acordo com a característica extraída (histograma de cor ou perfil de linha). A semelhança é baseada em um limiar $\tau$, o mesmo utilizado para determinar o número de grupos. Caso o valor medido seja menor que $\tau$, então o quadro-chave é eliminado. Em seguida, os quadros-chave resultantes são dispostos em ordem cronológica para facilitar o entendimento visual do resumo.

## 4.2 Avaliação dos Resumos

Os métodos de sumarização automática de vídeos devem ser avaliados para verificar a relevância dos quadros-chave selecionados. Apesar do problema de sumarização esteja sendo intensamente investigado, não existe um padrão ou um método ideal para avaliar a qualidade dos resumos produzidos (Money e Agius, 2008).

Neste trabalho, a avaliação é realizada de forma subjetiva, isto é, através de usuários. Duas formas de avaliação são propostas: pontuação média de opinião e comparação com os resumos dos usuários. Estas formas são detalhadas a seguir.

### 4.2.1 Pontuação Média de Opinião (PMO)

Nesta maneira de avaliação, a qualidade dos resumos é medida através da pontuação média de opinião, que corresponde a opinião ou grau de satisfação médio de um conjunto definido de pessoas quanto à qualidade do resumo produzido por um sistema de sumarização de vídeo, dentro de uma determinada escala, que varia de 1 a 5 (1 = “ruim”, 2 = “pobre”, 3 = “satisfatório”, 4 = “bom”, 5 = “excelente”).

Diferentemente da maioria das abordagens propostas, nas quais os usuários avaliam o conjunto de quadros-chave, tais como (Furini et al., 2007; Wang et al., 2007); na avaliação proposta, cada quadro-chave é avaliado separadamente, isto é, a relevância de cada quadro-chave é medida independentemente dos outros quadros-chave que compõem o resumo. Assim, objetiva-se obter uma avaliação quantitativa da qualidade, viabilizando a comparação direta entre diferentes métodos.

Na Figura 4.4, é ilustrada a avaliação proposta para a sumarização de vídeo. No passo 1, diferentes resumos são gerados por diversas técnicas para um mesmo vídeo. Em seguida

(passo 2), todos os quadros-chave que compõem os resumos são apresentados aos usuários. Estes devem atribuir um valor, de acordo com a escala definida, para cada quadro-chave. Este valor representa a opinião do usuário quanto à relevância do quadro-chave na identificação do conteúdo do vídeo. É importante ressaltar que a atribuição dos valores é feita sem a distinção entre os resumos, isto é, o usuário avalia cada quadro-chave sem saber a qual resumo o quadro-chave pertence. E por fim (passo 3), a pontuação média para cada resumo é computada através da seguinte fórmula:

$$PMO = \frac{soma\ da\ pontuacão\ dos\ quadros{-}chave}{número\ total\ de\ quadros{-}chave} \tag{4.1}$$

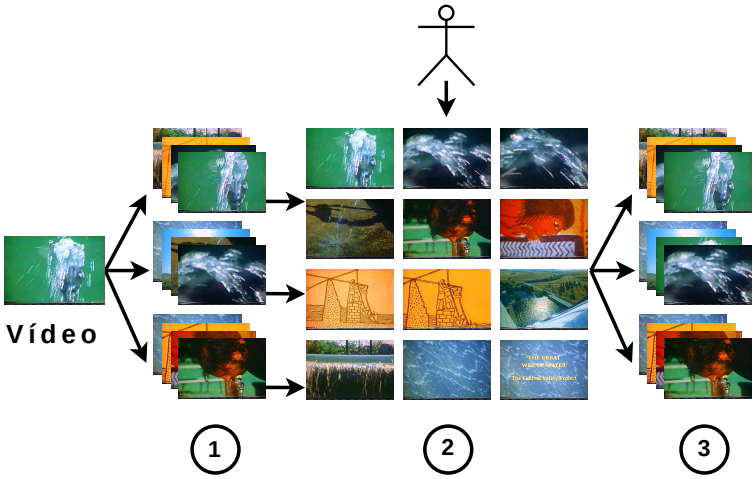



Figura 4.4: Diagrama do método de avaliação de resumos através da pontuação média de opinião.

### 4.2.2 Comparação com os Resumos dos Usuários (CRU)

Neste modo de avaliação, os resumos dos vídeos são criados manualmente pelos usuários e são comparados com os resumos produzidos automaticamente. Deste modo, o resumo do usuário é tomado como referência. O objetivo deste método é reduzir a subjetividade presente na tarefa de avaliação de resumos e obter comparações dos resultados de forma mais realista.

Esta forma de avaliação é similar ao método proposto por Guironnet et al. (2007), no qual a partir dos resumos criados pelos usuários, um resumo "ótimo" é construído automaticamente e este é comparado com os resumos gerados por diferentes técnicas. No entanto, esta construção não corresponde ao verdadeiro resumo do usuário, pois os resumos de $n$ usuários são condensados em um único resumo, o resumo "ótimo". Conseqüentemente, certos quadros-chave que um determinado usuário colocou no seu resumo podem não pertencer ao resumo "ótimo", como também o contrário pode acontecer.

Diferentemente, a avaliação proposta compara cada resumo do usuário com os resumos produzidos automaticamente, assim mantendo intacta a opinião do usuário. A comparação

é realizada quadro a quadro, comparando os histogramas de cor no espaço HSV, utilizando apenas o canal H. A distância entre os histogramas é medida através da distância de Manhattan (Equação 2.2). Dois quadros são considerados semelhantes, se a distância entre eles for menor que um limiar pré-determinado $\delta$. O valor do limiar utilizado, obtido por meio de testes experimentais, foi $0,5$.

Na Figura 4.5, é ilustrado o método de avaliação proposto. No passo 1, os quadros do vídeo (uma subamostragem) são apresentados aos usuários. Estes devem selecionar os quadros que, em sua opinião, conseguem representar o conteúdo do vídeo de forma concisa. Não há limite mínimo e nem limite máximo de quadros que podem ser selecionados. Em seguida (passo 2), os resumos dos usuários são comparados com os resumos produzidos automaticamente. A qualidade do resumo produzido (passo 3) é dada pela taxa de acerto $\text{CRU}_{acerto}$ e pela taxa de erro $\text{CRU}_{erro}$, descritas através das equações a seguir.

$$CRU_{acerto} = \frac{n_{CA}}{n_{RU}}, \tag{4.2}$$

$$CRU_{erro} = \frac{n_{NCA}}{n_{RU}}, \tag{4.3}$$

onde $n_{CA}$ corresponde ao número de quadros-chave casados do resumo automático, $n_{NCA}$ corresponde ao número de quadros-chave não-casados do resumo automático e $n_{RU}$ corresponde ao número total de quadros-chave do resumo do usuário.

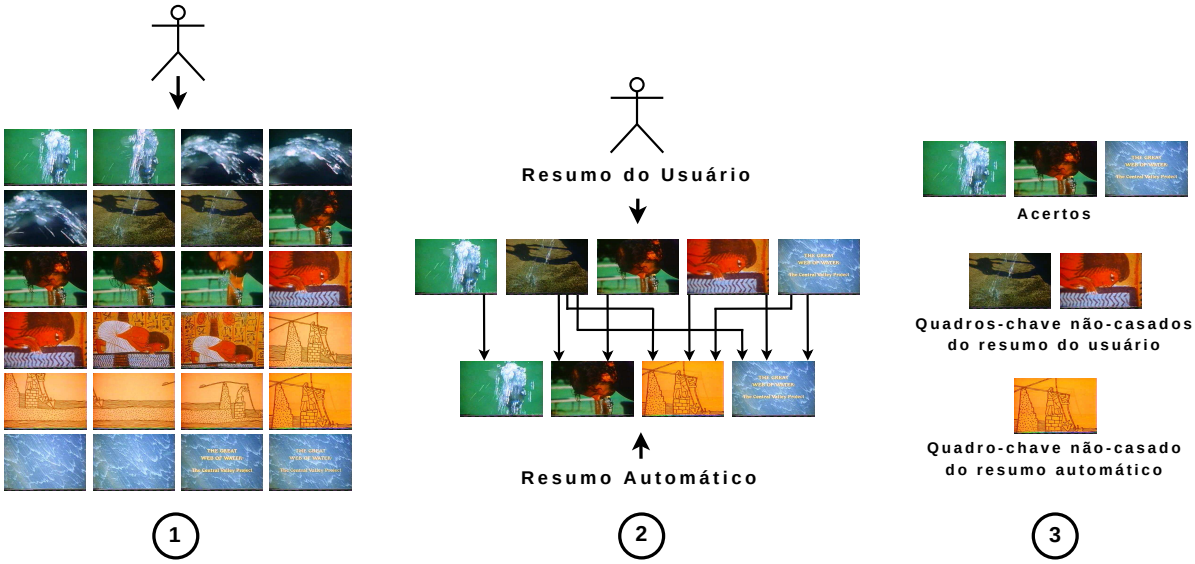



Figura 4.5: Diagrama do método de avaliação de resumos através da comparação com os resumos dos usuários.

O valor de $\text{CRU}_{acerto}$ varia de 0 (nenhum quadro-chave do resumo automático casa com

os quadros-chave do resumo do usuário, ou vice-versa) a 1 (todos os quadros-chave do resumo do usuário casam com os quadros-chave do resumo automático). Observe que $\text{CRU}_{acerto} = 1$ não significa necessariamente que todos os quadros-chave dos resumos (automático e usuário) são casados. Ou seja, se $n_{RU} < n_{RA}$ ($n_{RA}$ corresponde ao número total de quadros-chave do resumo automático), para $\text{CRU}_{acerto} = 1$, então determinados quadros-chave do resumo automático não serão casados.

No caso de $\text{CRU}_{erro}$, o valor varia de 0 (todos os quadros-chave do resumo automático casam com os quadros-chave do resumo do usuário) a $n_{RA}/n_{RU}$ (nenhum quadro-chave do resumo automático casa com os quadros-chave do resumo do usuário, ou vice-versa).

A qualidade mais alta é obtida quando $\text{CRU}_{acerto} = 1$ e $\text{CRU}_{erro} = 0$, ou seja, todos os quadros-chave do resumo automático casam com todos os quadros-chave do resumo do usuário.

## 4.3 Considerações

Neste capítulo, foram apresentadas a metodologia proposta para produzir resumos estáticos de vídeos e as formas de avaliação empregadas para medir a qualidade dos resumos produzidos. Como pode ser visto no próximo capítulo, os experimentos realizados durante a elaboração deste trabalho mostraram que a abordagem proposta provê resultados com qualidade superior às demais abordagens encontradas na literatura.

# Capítulo 5

# Resultados Experimentais

Neste capítulo, são apresentados os experimentos realizados e os resultados obtidos pela abordagem proposta. O capítulo está organizado da seguinte forma: na Seção 5.1, é descrita a base de dados utilizada, o repositório de vídeos Open Video; na Seção 5.2, é apresentada a ferramenta *ffmpeg*, a qual foi utilizada para separar o vídeo em quadros; e por fim, na Seção 5.3, são mostrados os experimentos realizados e são analisados os resultados alcançados.

## 5.1 Open Video

O repositório de vídeos Open Video[1] (OV) foi desenvolvido na Universidade da Carolina do Norte em Chapel Hill. O OV começou em 1998 com a digitalização de cerca de 195 vídeos. Logo em seguida, mais vídeos foram adicionados no repositório. A maioria dos vídeos foram disponibilizados pela Biblioteca de Vídeo Digital Informedia[2] (*Informedia Digital Video Library*), pelo Instituto Médico Howard Hughes[3] (*Howard Hughes Medical Institute*), pelo Arquivo de Prelinger[4] (*Prelinger Archive*) e pelas agências do governo americano, tais como a Administração Nacional de Arquivos e Registros[5] (*National Records and Archives Administration*) e a NASA[6] (sigla em inglês de *National Aeronautics and Space Administration*).

Atualmente, o repositório possui cerca de 4000 vídeos, que estão classificados de acordo com o formato (MPEG–1, MPEG–2, MPEG–4, QuickTime, RealMedia), o gênero (documentário, educacional, efêmero, histórico, vídeo-aula, entre outros), o tempo total de duração (menos de 1 minuto, 1 a 2 minutos, 2 a 5 minutos, 5 a 10 minutos e mais de 10 minutos), a cor (colorido ou preto e branco) e o áudio (com áudio ou mudo). O OV também disponibiliza os resumos (estáticos e dinâmicos) dos vídeos, os quais foram obtidos a partir da aplicação do algoritmo desenvolvido por (DeMenthon et al., 1998) e algumas intervenções manuais para refinar os resultados.

[1] http://www.open-video.org
[2] http://www.informedia.cs.cmu.edu
[3] http://www.hhmi.org
[4] http://www.archive.org/details/prelinger
[5] http://www.archives.gov
[6] http://www.nasa.gov

No presente trabalho, todos os vídeos utilizados estão no formato MPEG–1, com resolução 352 × 240 *pixels* a 30 quadros por segundo, abrangem diversos gêneros (documentário, educacional, efêmero e vídeo-aula), com tempo total de duração variando entre menos de 1 minuto e aproximadamente 4 minutos, são coloridos e possuem áudio.

A avaliação final da metodologia proposta foi feita em 50 vídeos do OV. Estes vídeos são os mesmos utilizados nos trabalhos desenvolvidos por Mundur et al. (2006) e Furini et al. (2007), nos quais os resumos gerados por cada técnica estão disponíveis. Assim, os resumos produzidos pela abordagem proposta são comparados com os resumos criados pelos autores supracitados e com os resumos do OV. Além disso, experimentos preliminares foram realizados em 20 outros vídeos do OV, selecionados aleatoriamente, a fim de investigar como os parâmetros (taxa de amostragem, número de grupos ($k$), número de cores quantizadas e intervalo entre perfis de linha) influenciam no desempenho e na qualidade dos resumos produzidos pelo sistema de sumarização de vídeo proposto.

Nas Tabelas 5.1 e 5.2, são listadas as informações dos vídeos utilizados nos experimentos. Estes são descritos através do código, do nome, do número total de quadros e do tempo total de duração do vídeo, sendo que o campo nome das tabelas é composto por duas partes: o rótulo dado ao vídeo para facilitar na visualização das próximas tabelas e o nome original do vídeo.

Tabela 5.1: Vídeos utilizados nos experimentos preliminares.

| **Código** | **Nome** | **#Quadros** | **Duração** |
|---|---|---|---|
| 347 | *v01 NASA 25th Anniversary Show, Segment 02* | 2.494 | 1:23 |
| 348 | *v02 NASA 25th Anniversary Show, Segment 03* | 4.267 | 2:22 |
| 349 | *v03 NASA 25th Anniversary Show, Segment 04* | 3.895 | 2:10 |
| 353 | *v04 NASA 25th Anniversary Show, Segment 08* | 2.775 | 1:32 |
| 6003 | *v05 Hurricanes and Computer Simulation* | 6.099 | 3:23 |
| 6024 | *v06 Drag Activity Part One* | 3.620 | 2:00 |
| 6059 | *v07 Model Testing* | 3.534 | 1:58 |
| 6079 | *v08 Computer Simulation* | 6.902 | 3:50 |
| 6116 | *v09 Aerosol Measurement and Remote Sensing* | 3.630 | 2:01 |
| 6121 | *v10 SAGE II and Picasso-Cena* | 3.609 | 2:01 |
| 6123 | *v11 Aerodynamic Forces* | 3.302 | 1:50 |
| 6127 | *v12 Wind Tunnels* | 4.662 | 2:35 |
| 6177 | *v13 Immune System* | 5.874 | 3:16 |
| 6189 | *v14 Flying a Plane* | 3.458 | 1:55 |
| 6345 | *v15 Flight Pioneers* | 6.019 | 3:20 |
| 6381 | *v16 Astronauts in Space* | 3.269 | 1:49 |
| 6391 | *v17 Space Suits* | 4.273 | 2:22 |
| 6394 | *v18 The Red Planet* | 4.306 | 2:23 |
| 6416 | *v19 Moon Phases* | 6.449 | 3:35 |
| 6559 | *v20 Oil Clean Up* | 3.537 | 1:58 |

Tabela 5.2: Vídeos utilizados nos experimentos.

| **Código** | **Nome** | **#Quadros** | **Duração** |
|---|---|---|---|
| 614 | *v21 The Great Web of Water, segment 01* | 3.279 | 1:50 |
| 615 | *v22 The Great Web of Water, segment 02* | 2.118 | 1:11 |
| 670 | *v23 The Great Web of Water, segment 07* | 1.745 | 0:59 |
| 684 | *v24 A New Horizon, segment 01* | 1.806 | 1:01 |
| 685 | *v25 A New Horizon, segment 02* | 1.797 | 1:00 |
| 686 | *v26 A New Horizon, segment 03* | 6.249 | 3:29 |
| 687 | *v27 A New Horizon, segment 04* | 3.192 | 1:47 |
| 689 | *v28 A New Horizon, segment 05* | 3.561 | 1:59 |
| 690 | *v29 A New Horizon, segment 06* | 1.944 | 1:05 |
| 661 | *v30 A New Horizon, segment 08* | 1.815 | 1:01 |
| 663 | *v31 A New Horizon, segment 10* | 2.517 | 1:24 |
| 635 | *v32 Take Pride in America, segment 01* | 2.691 | 1:30 |
| 637 | *v33 Take Pride in America, segment 03* | 3.261 | 1:49 |
| 4586 | *v34 Digital Jewelry: Wearable Technology for Every Day Life* | 4.204 | 3:00 |
| 4923 | *v35 HCIL Symposium 2002 – Introduction, segment 01* | 2.336 | 1:18 |
| 422 | *v36 Senses And Sensitivity, Introduct. to Lecture 1 presenter* | 4.221 | 2:20 |
| 425 | *v37 Senses And Sensitivity, Introduct. to Lecture 2* | 3.411 | 1:53 |
| 430 | *v38 Senses And Sensitivity, Introduct. to Lecture 3 presenter* | 4.566 | 2:32 |
| 538 | *v39 Senses And Sensitivity, Introduct. to Lecture 4 presenter* | 5.249 | 2:55 |
| 719 | *v40 Exotic Terrane, segment 01* | 2.940 | 1:38 |
| 720 | *v41 Exotic Terrane, segment 02* | 2.776 | 1:32 |
| 721 | *v42 Exotic Terrane, segment 03* | 2.676 | 1:29 |
| 722 | *v43 Exotic Terrane, segment 04* | 4.797 | 2:40 |
| 724 | *v44 Exotic Terrane, segment 06* | 2.425 | 1:21 |
| 726 | *v45 Exotic Terrane, segment 08* | 2.428 | 1:21 |
| 731 | *v46 America's New Frontier, segment 01* | 3.591 | 1:59 |
| 733 | *v47 America's New Frontier, segment 03* | 2.166 | 1:12 |
| 734 | *v48 America's New Frontier, segment 04* | 3.705 | 2:03 |
| 737 | *v49 America's New Frontier, segment 07* | 3.615 | 2:00 |
| 740 | *v50 America's New Frontier, segment 10* | 4.830 | 2:41 |
| 744 | *v51 The Future of Energy Gases, segment 03* | 2.934 | 1:37 |
| 746 | *v52 The Future of Energy Gases, segment 05* | 3.615 | 2:00 |
| 750 | *v53 The Future of Energy Gases, segment 09* | 1.884 | 1:02 |
| 753 | *v54 The Future of Energy Gases, segment 12* | 2.886 | 1:36 |
| 785 | *v55 Oceanfloor Legacy, segment 01* | 1.740 | 0:58 |
| 786 | *v56 Oceanfloor Legacy, segment 02* | 2.325 | 1:17 |
| 788 | *v57 Oceanfloor Legacy, segment 04* | 3.450 | 1:55 |
| 792 | *v58 Oceanfloor Legacy, segment 08* | 3.186 | 1:46 |
| 793 | *v59 Oceanfloor Legacy, segment 09* | 2.106 | 1:10 |
| 803 | *v60 The Voyage of the Lee, segment 05* | 2.094 | 1:09 |
| 813 | *v61 The Voyage of the Lee, segment 15* | 2.277 | 1:15 |
| 814 | *v62 The Voyage of the Lee, segment 16* | 2.619 | 1:27 |
| 838 | *v63 Hurricane Force – A Coastal Perspective, segment 03* | 2.310 | 1:17 |
| 839 | *v64 Hurricane Force – A Coastal Perspective, segment 04* | 5.310 | 2:57 |
| 850 | *v65 Drift Ice as a Geologic Agent, segment 03* | 2.742 | 1:31 |
| 852 | *v66 Drift Ice as a Geologic Agent, segment 05* | 2.187 | 1:12 |
| 853 | *v67 Drift Ice as a Geologic Agent, segment 06* | 2.425 | 1:30 |
| 854 | *v68 Drift Ice as a Geologic Agent, segment 07* | 1.950 | 1:05 |
| 855 | *v69 Drift Ice as a Geologic Agent, segment 08* | 3.618 | 2:00 |
| 857 | *v70 Drift Ice as a Geologic Agent, segment 10* | 1.407 | 0:46 |

## 5.2 Ferramenta *ffmpeg*

Para a realização da etapa de separação do vídeo em quadros foi utilizada a ferramenta *ffmpeg*, que é parte do projeto FFmpeg[7]. Este projeto foi desenvolvido na plataforma GNU/Linux, mas pode ser compilado em outros sistemas operacionais (Apple Mac OS X, Microsoft Windows e AmigaOS).

A ferramenta *ffmpeg*, implementada em linguagem C, permite fazer a extração dos quadros do vídeo de forma simples, bastante rápida e via linha de comando. Além disso, permite fazer a conversão de arquivos de vídeo e áudio em diversos formatos, a conversão de taxas de amostragem e o redimensionamento do vídeo.

Esta ferramenta foi utilizada neste trabalho por se tratar de um *software* livre e principalmente por realizar de maneira rápida a extração dos quadros do vídeo.

## 5.3 Experimentos

A apresentação dos experimentos está dividida em duas partes: primeiro, são mostrados os experimentos preliminares, os quais foram efetuados com os vídeos listados na Tabela 5.1; e segundo, a partir da análise obtida dos experimentos preliminares, novos experimentos são executados utilizando os vídeos descritos na Tabela 5.2.

Os experimentos foram conduzidos em uma máquina com processador Intel® Core™ Duo 1.83 GHz, 2GB de memória RAM, disco rígido Fujitsu com interface SATA-150 e rotação de 5400 RPM, e sistema operacional Microsoft Windows XP. Para implementação computacional foi utilizada a linguagem de programação C++ e o ambiente de programação Microsoft Visual Studio 2005.

### 5.3.1 Experimentos Preliminares

Nestes experimentos, a metodologia proposta para o processo de sumarização de vídeo é aplicada de forma simplificada com o propósito de facilitar a análise dos parâmetros do algoritmo, identificar possíveis problemas no método proposto e apontar outras soluções. A metodologia simplificada é composta pelos passos descritos a seguir.

**Passo 1** Segmentação do vídeo em quadros. Estes são obtidos conforme a taxa de amostragem utilizada. Isto é, o número total de quadros extraídos do vídeo corresponde ao número de quadros dividido pela taxa de amostragem aplicada. Diferentes taxas foram analisadas.

**Passo 2** Extração das características de cor presentes nos quadros. Nesta etapa são aplicados os métodos de histograma de cor e o perfil de linha (horizontal, vertical e diagonal).

**Passo 3** Agrupamento dos quadros por meio do algoritmo *k-means*. A medida de similaridade utilizada é a distância Euclidiana. Nesta etapa, os experimentos foram executados

[7] http://ffmpeg.mplayerhq.hu

para diferentes valores de $k$ (número de grupos).

**Passo 4** Identificação dos quadros-chave. Para cada grupo, um quadro-chave é identificado.

**Passo 5** Eliminação dos quadros-chave semelhantes e disposição dos quadros-chave restantes em ordem cronológica para facilitar o entendimento visual do resultado. A qualidade dos resumos é estimada através da PMO (Pontuação Média de Opinião) (veja a Seção 4.2.1).

### Análise da Taxa de Amostragem

Para analisar a taxa de amostragem, os experimentos foram realizados em oito vídeos escolhidos aleatoriamente da amostra de 20 vídeos. Para cada taxa utilizada, a qualidade do resumo e o tempo necessário para produzi-lo foram medidos com a finalidade de determinar quantos quadros do vídeo devem ser analisados. As taxas foram definidas de acordo com os gêneros dos vídeos utilizados nesses experimentos preliminares (documentário e educacional), os quais são compostos por tomadas longas.

Na Tabela 5.3, são apresentados os tempos de execução para a geração dos resumos utilizando todos os quadros do vídeo, um quadro a cada 30, 45, 60, 75 e 90 quadros. O tempo necessário para segmentar os vídeos em quadros não está incluso na Tabela 5.3. Para criar os resumos, a extração de características dos quadros dos vídeos foi obtida através do histograma de cor (no espaço de cor RGB com 256 cores em cada canal) e o número de grupos foi fixado em 15 grupos, isto é, $k = 15$.

Tabela 5.3: Tempo de execução em segundos para a sumarização de vídeos utilizando todos os quadros do vídeo e as taxas de amostragem um quadro a cada 30, 45, 60, 75 e 90 quadros.

| **Vídeos** | | **Tempo de Execução (s)** | | | | | |
|---|---|---|---|---|---|---|---|
| **Nome** | **Duração** | **1** | **30** | **45** | **60** | **75** | **90** |
| *v02* | 2:22 | 96,39 | 3,15 | 2,17 | 1,68 | 1,28 | 1,21 |
| *v03* | 2:10 | 83,86 | 2,93 | 1,98 | 1,53 | 1,20 | 1,08 |
| *v04* | 1:32 | 59,11 | 2,07 | 1,43 | 1,12 | 0,90 | 0,79 |
| *v05* | 3:23 | 160,53 | 4,87 | 3,44 | 2,58 | 2,09 | 1,81 |
| *v06* | 2:00 | 87,34 | 2,95 | 2,00 | 1,55 | 1,27 | 1,13 |
| *v08* | 3:50 | 172,85 | 5,43 | 3,80 | 3,01 | 2,26 | 2,00 |
| *v10* | 2:01 | 85,76 | 2,92 | 1,43 | 1,62 | 1,26 | 1,09 |
| *v15* | 3:20 | 162,08 | 4,71 | 3,47 | 2,51 | 2,04 | 1,72 |
| **Média** | | 113,49 | 3,63 | 2,47 | 1,95 | 1,54 | 1,35 |

A partir dos experimentos realizados, foi observado que a qualidade dos resumos produzidos não mostrou nenhuma diferença significativa, exceto para a taxa de um quadro a cada 90 quadros. Além disso, pode-se notar que quanto menor a taxa de amostragem, menor foi o tempo de execução do algoritmo de sumarização. Como o método apresentou, em média, o menor tempo de execução aplicando a taxa de um quadro a cada 75 quadros[8] em vez das

[8] A taxa de amostragem de um quadro a cada 90 quadros foi desprezada porque influenciou negativamente na qualidade dos resumos.

outras taxas de amostragem, então foi feita uma comparação estatística entre esta taxa e as demais. Pelos dados da Tabela 5.4, pode-se afirmar, com 95% de confiança, que o algoritmo de sumarização apresenta um tempo de execução menor empregando a taxa de amostragem de um quadro a cada 75 quadros, uma vez que em cada comparação o intervalo de confiança não incluiu o valor zero (Jain, 1992). Sendo assim, os próximos experimentos foram realizados utilizando esta taxa.

Tabela 5.4: Comparação entre tempos de execução em segundos para as diferentes taxas de amostragem (um quadro a cada 75 quadros e as demais taxas). $T_{75}$ corresponde a um quadro a cada 75 quadros, $T_1$ utiliza todos os quadros, $T_{30}$ um quadro a cada 30 quadros, $T_{45}$ um quadro a cada 45 quadros e $T_{60}$ um quadro a cada 60 quadros.

| **Diferença** | **Média** | **Desvio Padrão** | **Intervalo de Confiança (95%) mínimo** | **máximo** |
|---|---|---|---|---|
| $T_{75} - T_1$ | –111,95 | 43,70 | –148,49 | –75,42 |
| $T_{75} - T_{30}$ | –2,09 | 0,69 | –2,67 | –1,51 |
| $T_{75} - T_{45}$ | –0,93 | 0,48 | –1,33 | –0,53 |
| $T_{75} - T_{60}$ | –0,41 | 0,16 | –0,55 | –0,28 |

**Análise das Técnicas de Extração de Características e Avaliação dos Resumos através da PMO**

Para analisar as técnicas de extração de características (histograma de cor e perfil de linha) foi aplicado um projeto fatorial $2^p$ (Jain, 1992), com $p = 2$ ou $p = 3$ se a técnica utilizada foi histograma de cor ou perfil de linha, respectivamente. Para simplificar o modelo experimental, os fatores que afetam o desempenho do processo de sumarização de vídeo foram representados em dois níveis, como mostrado a seguir:

- Número de grupos ($k$): 15 ou 35 grupos.
- Número de cores quantizadas ($n_c$): 16 ou 256 cores.
- Intervalo entre perfis de linha ($i_p$): 10 ou 40 linhas.

Para estes experimentos, são utilizados os mesmos vídeos do experimento anterior. Os tempos de execução para a produção dos resumos são mostrados na Tabela 5.5 e na Tabela 5.6, onde a configuração do 5º experimento da Tabela 5.6 apresentou o menor tempo de execução para o processo de sumarização.

A comparação do desempenho do algoritmo, para as diferentes técnicas de extração de características (histograma de cor $H$, perfil de linha horizontal $P_H$, perfil de linha vertical $P_V$ e perfil de linha diagonal $P_D$), foi feita para o menor tempo de execução obtido por cada técnica. Com 95% de confiança (veja a Tabela 5.7), pode-se afirmar que o algoritmo de sumarização é executado em um tempo menor utilizando os perfis de linha do que o histograma. Além disso, para o mesmo nível de confiança, os resultados obtidos utilizando os perfis de linha são significativamente diferentes, pois os intervalos de confiança não incluem o valor zero.

Tabela 5.5: Tempo de execução para a sumarização de vídeos utilizando histograma.

| $k$ | $n_c$ | **Tempo de Execução (s)** |
|---|---|---|
| 15 | 16 | **1,48** |
| 35 | 16 | 1,56 |
| 15 | 256 | 1,66 |
| 35 | 256 | 1,91 |

Tabela 5.6: Tempo de execução para a sumarização de vídeos utilizando perfis de linha.

| $k$ | $n_c$ | $i_p$ | **Tempo de Execução (s)** | | |
|---|---|---|---|---|---|
| | | | **Horizontal** | **Vertical** | **Diagonal** |
| 15 | 16 | 10 | 0,83 | 0,97 | 1,07 |
| 35 | 16 | 10 | 1,23 | 1,24 | 1,26 |
| 15 | 256 | 10 | 0,81 | 0,95 | 1,12 |
| 35 | 256 | 10 | 0,96 | 1,13 | 1,45 |
| 15 | 16 | 40 | **0,62** | 0,87 | 0,82 |
| 35 | 16 | 40 | 0,67 | 1,02 | 1,02 |
| 15 | 256 | 40 | 0,86 | **0,72** | **0,79** |
| 35 | 256 | 40 | 1,11 | 0,78 | 0,90 |

Tabela 5.7: Comparação entre os tempos de execução para as diferentes técnicas de extração de características.

| **Diferença** | **Média** | **Desvio Padrão** | **Intervalo de Confiança (95%)** | |
|---|---|---|---|---|
| | | | **mínimo** | **máximo** |
| $P_H - H$ | –0,88 | 0,31 | –1,14 | –0,62 |
| $P_V - H$ | –0,78 | 0,29 | –1,02 | –0,54 |
| $P_D - H$ | –0,70 | 0,26 | –0,92 | –0,48 |
| $P_H - P_V$ | –0,10 | 0,04 | –0,14 | –0,07 |
| $P_H - P_D$ | –0,18 | 0,09 | –0,26 | –0,11 |
| $P_V - P_D$ | –0,08 | 0,08 | –0,15 | –0,01 |

Contudo, como a qualidade do resumo é mais importante do que o tempo necessário para produzi-lo, então os resumos gerados para os experimentos que apresentaram o melhor desempenho, utilizando histograma e perfil de linha, foram avaliados por 10 pessoas objetivando-se medir a qualidade destes. Esta mensuração foi obtida através da PMO (Equação 4.1). Antes de avaliar o resumo, as pessoas podiam assistir ao vídeo quantas vezes fosse necessário para compreender o seu conteúdo. Na Tabela 5.8, é apresentada a PMO dos resumos para cada vídeo.

O resultado da avaliação dos usuários mostrou que o uso do perfil de linha horizontal para a sumarização de vídeo fornece resultados com qualidade superior em relação às demais técnicas utilizadas. Para validar esta afirmação, foi feita a comparação entre a qualidade alcançada utilizando o perfil de linha horizontal e as outras técnicas. Na Tabela 5.9, são mostrados

Tabela 5.8: Resultado da avaliação dos usuários através da PMO.

| Vídeos | PMO | | | |
|---|---|---|---|---|
| | Histog. | Perfil de Linha | | |
| | | Horiz. | Vert. | Diag. |
| *v02* | 3,3 | 3,4 | 3,5 | 3,3 |
| *v03* | 3,3 | 3,7 | 3,5 | 3,3 |
| *v04* | 3,6 | 3,7 | 3,5 | 3,7 |
| *v05* | 3,4 | 3,5 | 3,4 | 3,3 |
| *v06* | 2,8 | 3,3 | 2,9 | 2,9 |
| *v08* | 3,3 | 3,3 | 3,2 | 3,1 |
| *v10* | 2,9 | 3,4 | 3,1 | 2,8 |
| *v15* | 3,7 | 3,4 | 3,3 | 3,1 |
| **Média** | 3,3 | **3,5** | 3,3 | 3,2 |

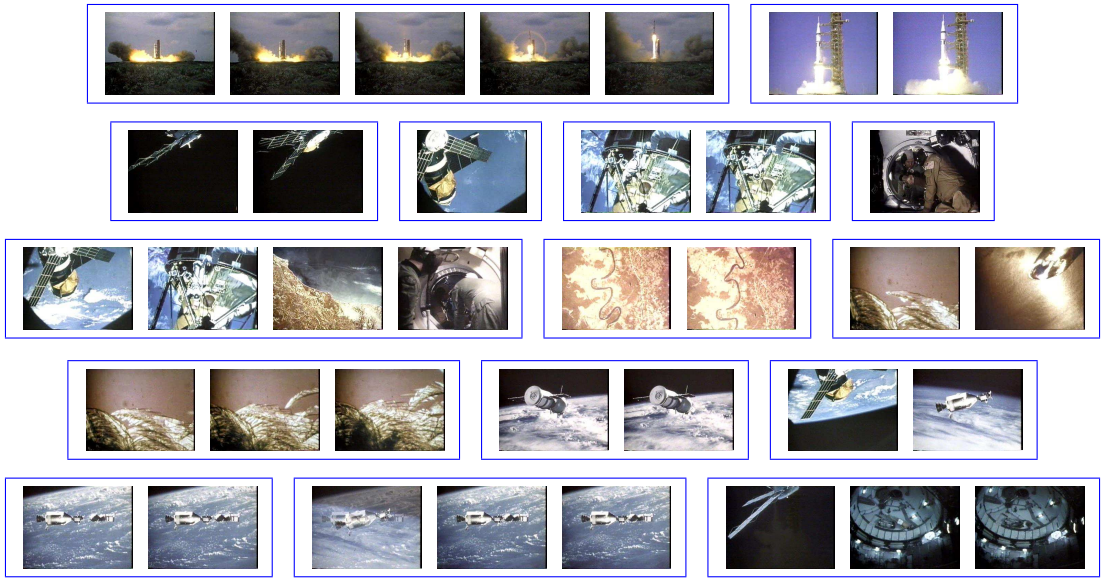
Tabela 5.9: Comparação entre as qualidades dos resumos para as diferentes técnicas de extração de características.

| Diferença | Média | Desvio Padrão | Intervalo de Confiança (95%) mínimo | máximo |
|---|---|---|---|---|
| $P_H - H$ | 0,18 | 0,26 | –0,04 | 0,39 |
| $P_H - P_V$ | 0,15 | 0,15 | 0,02 | 0,28 |
| $P_H - P_D$ | 0,28 | 0,18 | 0,13 | 0,42 |

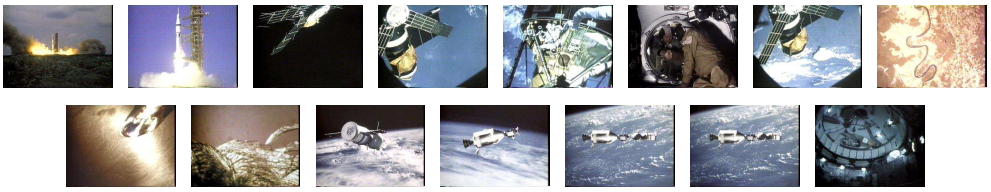
os resultados desta comparação. Com 95% de confiança, pode-se afirmar que os resumos apresentam qualidade superior utilizando o perfil de linha horizontal em vez do perfil de linha vertical ou diagonal. No entanto, para este mesmo nível de confiança, a qualidade dos resumos produzidos utilizando o perfil de linha horizontal ou o histograma é equiparável, ou seja, não se pode afirmar que a utilização do perfil de linha horizontal no processo de sumarização fornece resultados com qualidade superior em relação ao histograma, ou vice-versa.

Assim, como o perfil de linha horizontal e o histograma não apresentam diferença em relação à qualidade dos resumos, mas o perfil de linha horizontal produz os resumos em um menor tempo, a geração dos resumos dos 20 vídeos (veja a Tabela 5.1) foi feita utilizando a melhor configuração experimental obtida – 16 cores, intervalo de 40 linhas entre os perfis (ou seja, 6 perfis horizontais) e diferentes números de grupos (15, 20, 25, 30 e 35 grupos). Na Tabela 5.10, são apresentadas as médias do tempo de execução do algoritmo de sumarização para cada grupo e a PMO obtida. Estes resumos foram avaliados pelos mesmos 10 usuários.

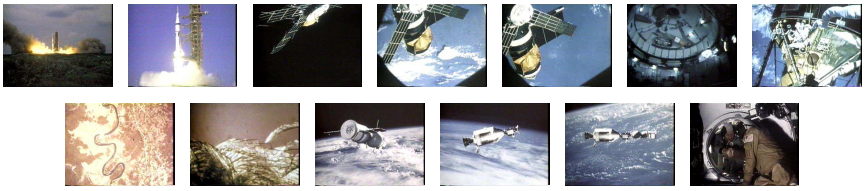
Para ilustrar o funcionamento do método proposto, os principais passos são representados na Figura 5.1. O resumo foi gerado para a melhor configuração experimental e $k = 15$. Primeiramente, os 15 grupos obtidos por meio do processo de agrupamento são ilustrados (Figura 5.1.a). Em seguida, os quadros-chave são selecionados, um quadro-chave para cada grupo (Figura 5.1.b). E por fim, os quadros-chave semelhantes são eliminados e os restantes são dispostos em ordem cronológica (Figura 5.1.c).



(a) Agrupamento dos quadros. Os retângulos representam os grupos formados.



(b) Seleção dos quadros-chave.



(c) Eliminação dos quadros-chave semelhantes.

Figura 5.1: Ilustração dos principais passos da metodologia proposta para o vídeo *NASA 25th Anniversary Show, Segment 08 (v04)*.

Tabela 5.10: Tempo de execução para a sumarização de vídeos utilizando o perfil de linha horizontal e resultado da avaliação dos usuários através da PMO.

| $k$ | **Tempo de Execução (s)** | **PMO** |
|---|---|---|
| 15 | 0,59 | 3,6 |
| 20 | 0,64 | 4,2 |
| 25 | 0,63 | **4,3** |
| 30 | 0,64 | 4,0 |
| 35 | 0,65 | 3,7 |

### Comparação com os Resumos do Open Video

Para os 20 vídeos listados na Tabela 5.1, foi realizada uma comparação entre os resumos produzidos nos experimentos preliminares e os resumos do OV. Para não favorecer ou prejudicar o método proposto, o número de quadros-chave do resumo gerado foi baseado no número de quadros-chave dos resumos do repositório de vídeos. Ou seja, se o resumo do OV contém cinco quadros-chave, então o número de quadros-chave para o método proposto também é fixado em cinco quadros-chave. Assim, o número máximo de quadros-chave será o mesmo para ambas abordagens.

Tabela 5.11: PMO e número de quadros-chave dos resumos gerados pelo Open Video e pelo método proposto.

| **Vídeos** | **PMO** | | **#Quadros-chave** | |
|---|---|---|---|---|
| | **Open Video** | **Método Proposto** | **Open Video** | **Método Proposto** |
| *v01* | 4,0 | **4,4** | 20 | 10 |
| *v02* | 3,6 | **3,8** | 26 | 10 |
| *v03* | 3,8 | 3,8 | 29 | 15 |
| *v04* | 3,8 | 3,8 | 14 | 9 |
| *v05* | 3,8 | 3,8 | 22 | 13 |
| *v06* | **3,4** | 3,3 | 6 | 6 |
| *v07* | 3,3 | 3,3 | 12 | 10 |
| *v08* | **3,8** | 3,6 | 19 | 11 |
| *v09* | 3,4 | **3,7** | 12 | 8 |
| *v10* | **3,8** | 3,7 | 12 | 7 |
| *v11* | 3,8 | **4,1** | 12 | 9 |
| *v12* | **3,7** | 3,6 | 12 | 10 |
| *v13* | 3,7 | **4,0** | 6 | 5 |
| *v14* | **3,9** | 3,5 | 7 | 7 |
| *v15* | 3,7 | **4,0** | 19 | 9 |
| *v16* | 3,5 | 3,5 | 18 | 10 |
| *v17* | 3,7 | **4,0** | 8 | 6 |
| *v18* | **3,8** | 3,5 | 10 | 6 |
| *v19* | 3,3 | **3,6** | 13 | 8 |
| *v20* | 3,0 | **3,6** | 12 | 8 |

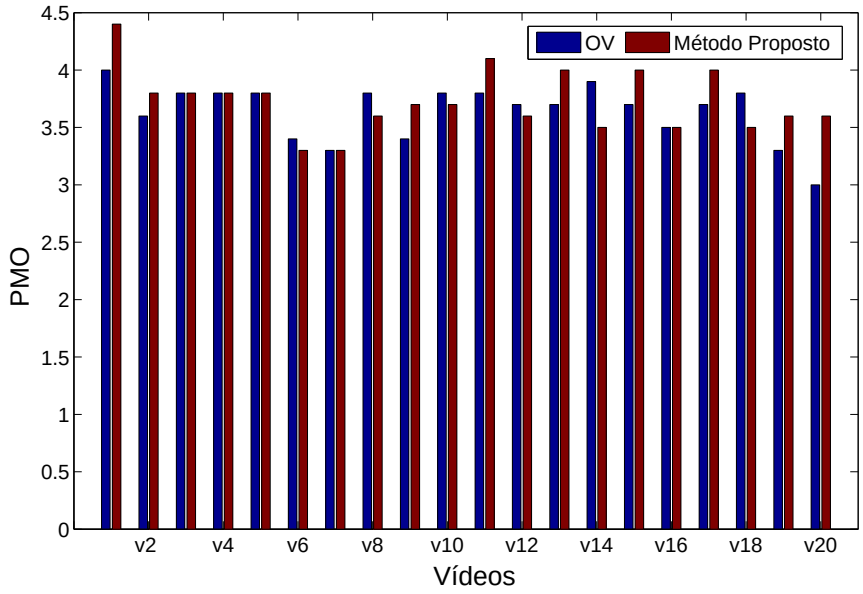



Figura 5.2: Gráfico da comparação da PMO obtida para cada vídeo.

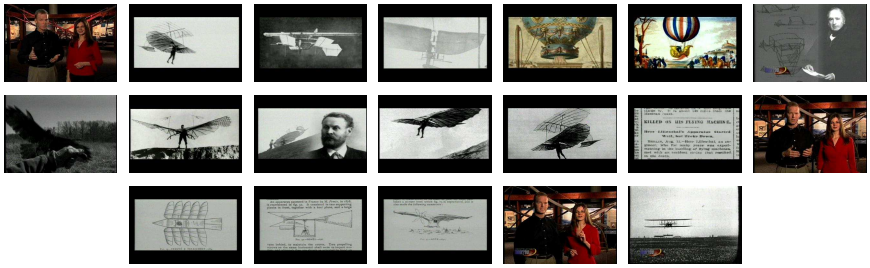
Na Tabela 5.11, são apresentados a qualidade e o número de quadros-chave de cada resumo. Na Figura 5.2, é representado o gráfico da comparação da PMO obtida para cada vídeo.

De acordo com os resultados experimentais obtidos, os resumos produzidos pelo método proposto apresentaram qualidade superior para nove vídeos dentre os 20 vídeos; Cinco resumos obtiveram a mesma PMO; E para seis vídeos, os resumos gerados pelo OV apresentaram qualidade superior. Pode-se notar ainda que o melhor resultado alcançado pelo método proposto apresentou uma PMO igual a 4,4 em contraste ao melhor resultado do OV que foi igual a 4,0. As piores PMOs das duas abordagens foram iguais a 3,3. Além disso, o método proposto apresentou cinco resumos com PMO maior ou igual a 4,0, enquanto que o OV obteve apenas uma PMO igual a 4,0.

Para confirmar que o método proposto fornece resumos com qualidade superior em relação aos resumos do OV, foi feita a comparação entre as PMOs obtidas por cada abordagem. Na Tabela 5.12, é mostrado o resultado desta comparação. Como o intervalo de confiança inclui o valor zero, então não se pode afirmar que, para um nível de 95% de confiança, os resumos gerados pelo método proposto apresentam qualidade superior em relação aos resumos do OV, ou vice-versa.

Para ilustrar esta comparação são mostrados os resultados para três vídeos. Na Figura 5.3, o resumo gerado pelo método proposto apresentou qualidade superior em relação ao resumo do OV. Por outro lado, na Figura 5.4, o resumo do OV mostrou um resultado melhor. E, na Figura 5.5, os resumos apresentaram a mesma PMO. Todos os resultados são mostrados no Apêndice A.

(a) Método proposto. PMO = 4,0.



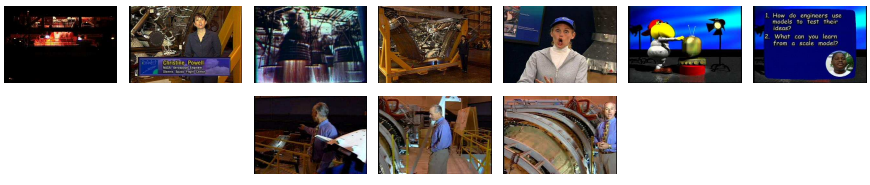
(b) Open Video. PMO = 3,7.

Figura 5.3: Resumos produzidos para o vídeo *Flight Pioneers (v15)*.

(a) Método proposto. PMO = 3,5.

(b) Open Video. PMO = 3,9.

Figura 5.4: Resumos produzidos para o vídeo *Flying a Plane (v14)*.



(a) Método proposto. PMO = 3,3.

(b) Open Video. PMO = 3,3.

Figura 5.5: Resumos produzidos para o vídeo *Model Testing (v07)*.

Tabela 5.12: Comparação entre as qualidades dos resumos produzidos pelo método proposto e pelo Open Video.

| Diferença | Média | Desvio Padrão | Intervalo de Confiança (95%) mínimo | máximo |
|---|---|---|---|---|
| Mét. Prop. – OV | 0,09 | 0,26 | –0,03 | 0,21 |

### Análise dos Resultados Preliminares

Conforme os resultados experimentais obtidos, observou-se a necessidade da melhoria de algumas etapas da metodologia empregada. Estas idéias, listadas a seguir, são aplicadas nos experimentos finais.

**Modelo de cor** O modelo RGB, aplicado nos experimentos preliminares, não é capaz de representar todas as cores perceptíveis pelo sistema visual humano, apesar de ser uma boa aproximação. Uma variação de luminosidade ou saturação no modelo RGB varia de forma irregular dentro do espaço de cor, sendo necessárias outras formas de representação para capturar as informações de cor mais de acordo com sistema visual humano. Diante destes motivos, verificou-se a necessidade de utilizar um modelo de cor mais robusto à variação de cor, como por exemplo o HSV, aplicado em (Zhuang et al., 1998; Mundur et al., 2006; Furini et al., 2007) e nos experimentos apresentados na seção seguinte.

**Pré-processamento dos quadros** O objetivo do pré-processamento proposto consiste em eliminar quadros inexpressivos, isto é, quadros completamente pretos (ou de outra cor), que geralmente ocorrem no vídeo devido a efeitos de *fade-in* ou *fade-out* ou devido ao uso de *flashes*.

**Taxa de amostragem** A taxa de amostragem aplicada nos experimentos preliminares (um quadro a cada 75 quadros) mostrou-se adequada para os gêneros dos vídeos utilizados (documentário e educacional), os quais são compostos por tomadas longas. Nos próximos experimentos, outros gêneros foram usados (tais como efêmero, vídeo-aula), além dos gêneros mencionados anteriormente. Como estes gêneros não são compostos por apenas tomadas longas, então a baixa taxa de amostragem empregada nos experimentos preliminares pode não ser adequada para os experimentos finais. Sendo assim, para garantir que não houvesse perda de informação na extração dos quadros do vídeo, a taxa de amostragem utilizada nos experimentos finais corresponde ao valor da taxa de quadros por segundo do vídeo.

**Extração de características** De acordo com os resultados obtidos, o histograma e o perfil horizontal não apresentaram diferença estatisticamente significativa em relação à qualidade dos resumos produzidos. Embora o tempo de execução do algoritmo seja menor utilizando o perfil de linha horizontal, os próximos experimentos são realizados aplicando o histograma de cor. Esta decisão foi tomada devido à diferença entre as dimensões dos

vetores de características das duas técnicas. Para o perfil de linha horizontal, a dimensão do vetor é dada pela largura do quadro do vídeo, enquanto que a dimensão do vetor, para o histograma de cor, é dada pelo número de *bins* de cada canal do modelo de cor, que pode ser reduzida por meio da quantização de cores, o que viabiliza a aplicação do algoritmo de agrupamento.

**Número de grupos ($k$)** Um problema evidente foi a determinação manual do valor de $k$ para a realização dos experimentos. Embora vários testes tenham sido realizados para diferentes valores de $k$, o mesmo valor de $k$ foi aplicado em vários vídeos, que apresentam características diferentes (tais como a duração do vídeo e o conteúdo) que não foram levadas em consideração. Para contornar este problema, foi proposto um método (apresentado na Seção 4.1.4) para determinar automaticamente o valor de $k$ para cada vídeo.

**Identificação dos quadros-chave** A solução mais comum é seleção do ponto mais próximo do centróide do grupo como o quadro-chave do grupo. A maioria dos trabalhos pesquisados na literatura seleciona um quadro-chave de cada grupo. Esta idéia foi aplicada neste trabalho e por diversos outros (Hanjalic e Zhang, 1999; Uchihashi et al., 1999; Gong e Liu, 2000; Yu et al., 2004; Hadi et al., 2006; Mundur et al., 2006; Furini et al., 2007). No entanto, ocorrem situações em que determinados grupos são formados por apenas um quadro (ou um número pequeno de quadros em relação aos outros grupos), enquanto que para o mesmo vídeo outros grupos são compostos de vários quadros. Uma estratégia para eliminar estes grupos de tamanho pequeno (ou seja, grupos que representam trechos muito curtos do vídeo) foi proposta por Zhuang et al. (1998), aplicada por Cámara-Chávez (2007) e adaptada neste trabalho, conforme descrita na Seção 4.1.5. A estratégia adaptada e a seleção de um quadro-chave para cada grupo são avaliadas nos experimentos finais.

**Avaliação dos resumos** No método de avaliação proposto (PMO), a relevância de cada quadro-chave é medida, através de uma escala, independentemente dos outros quadros-chave que compõem o resumo. Entretanto, a qualidade do resumo também depende da avaliação do conjunto de quadros-chave como um todo. Por exemplo, um quadro-chave avaliado separadamente pode não ser relevante, no entanto, este mesmo quadro-chave avaliado dentro do resumo (ou seja, considerando os outros quadros-chave que compõem o resumo) pode ser importante na representação do conteúdo do vídeo. Outro ponto observado foi a dificuldade relatada pelos usuários diante da forma de avaliação proposta, devido ao alto grau de subjetividade da tarefa de atribuição de valores para os quadros-chave. Para reduzir esta subjetividade e estimar a qualidade dos resumos de forma mais realista, é aplicada a comparação entre os resumos produzidos automaticamente e os resumos criados pelos usuários. Este modo de avaliação foi proposto por Guironnet et al. (2007) e adaptado neste trabalho, conforme apresentado na Seção 4.2.2. A qualidade dos resumos gerados nos experimentos finais é obtida através desta forma de avaliação.

### 5.3.2 Experimentos Finais

Nestes experimentos, a metodologia aplicada para o processo de sumarização de vídeo contém basicamente os mesmos passos da metodologia avaliada anteriormente. No entanto, algumas mudanças foram realizadas com o objetivo de solucionar os problemas identificados na metodologia anterior. A metodologia empregada nestes experimentos, denominada VSUMM (*Video SUMMarization*), é composta pelos passos descritos a seguir.

**Passo 1** Segmentação do vídeo em quadros. Estes são obtidos conforme a taxa de quadros por segundo do vídeo. Isto é, o número total de quadros extraídos do vídeo corresponde ao número de quadros dividido pelo valor da taxa de quadros por segundo do vídeo.

**Passo 2** Extração das características de cor presentes nos quadros utilizando o histograma de cor (no espaço de cor HSV, utilizando apenas o canal H e 16 cores). Nesta etapa, os quadros inexpressivos são eliminados.

**Passo 3** Agrupamento dos quadros através do algoritmo *k-means*. O valor de $k$ é determinado automaticamente para cada vídeo. Nesta etapa, diversas medidas de similaridade são avaliadas.

**Passo 4** Identificação dos quadros-chave. Esta identificação pode ocorrer de duas formas: 1) para cada grupo, um quadro-chave é selecionado; ou 2) para cada grupo-chave[9], um quadro-chave é identificado.

**Passo 5** Eliminação dos quadros-chave semelhantes e disposição dos quadros-chave restantes em ordem cronológica para facilitar o entendimento visual do resultado. A qualidade dos resumos é estimada através da CRU (Comparação com os Resumos dos Usuários) (veja a Seção 4.2.2).

#### Análise das Medidas de Similaridade

A medida de similaridade aplicada no processo de agrupamento é determinante para o bom resultado do processo, pois, se a medida utilizada não representa bem a distância entre dois quadros quaisquer, então quadros similares poderão ser agrupados em grupos diferentes, ou ainda, quadros não-similares poderão ser dispostos no mesmo grupo. Sendo assim, as medidas de similaridade listadas na Tabela 2.1 são testadas a fim de identificar a medida que consegue melhor mensurar a distância entre os quadros do vídeo.

Para analisar as medidas de similaridade, as distâncias par-a-par entre os vetores de características dos quadros de um determinado vídeo foram calculadas e normalizadas no intervalo [0,1]. Para facilitar a interpretação e análise destas distâncias, foram computados os seus percentis (10, 20, ..., 90). Na Figura 5.6, é mostrado o gráfico da distribuição de cada distância em percentis.

[9]Um grupo é considerado como *grupo-chave* se o seu tamanho é maior que $N/2k$, onde $N$ representa o número de quadros analisados do vídeo.

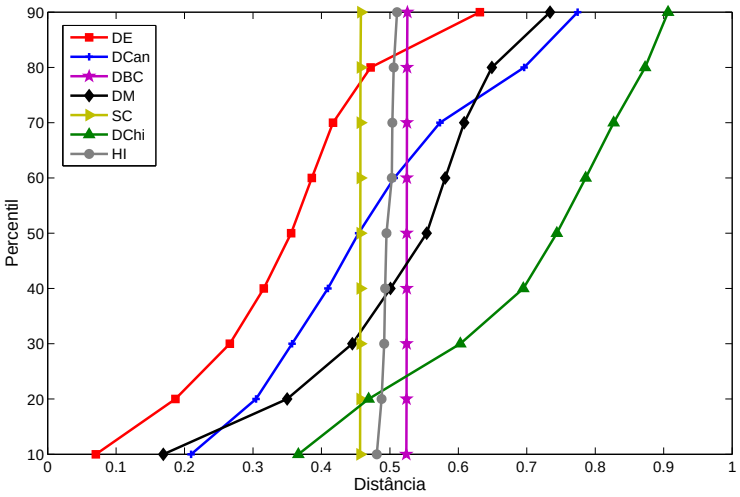



Figura 5.6: Gráfico da distribuição das distâncias, em percentis, entre os vetores de características dos quadros do vídeo *The Great Web of Water, segment 01*. $DE$ corresponde a distância Euclidiana, $DCan$ a distância de Canberra, $DBC$ a distância de Bray–Curtis, $DM$ a distância de Manhattan, $SC$ a similaridade do cosseno, $DChi$ a distância $\chi^2$ e $HI$ a intersecção de histogramas.

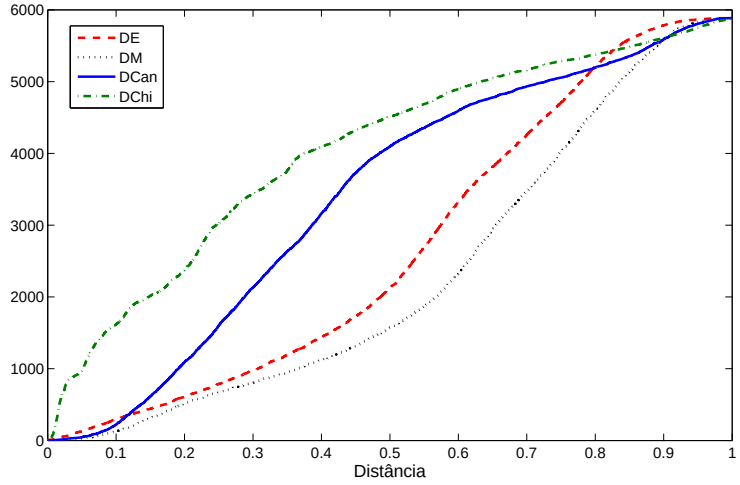



Figura 5.7: Gráfico das distâncias entre os vetores de características dos quadros do vídeo *The Great Web of Water, segment 01*. $DE$ corresponde a distância Euclidiana, $DM$ a distância de Manhattan, $DCan$ a distância de Canberra e $DChi$ a distância $\chi^2$.

De acordo com os resultados apresentados no gráfico, pode-se notar claramente que a distância de Bray–Curtis, a similaridade do cosseno e a intersecção de histogramas não conseguiram representar as distâncias entre os quadros do vídeo, pois as distâncias calculadas se concentraram praticamente em um mesmo valor. Por outro lado, a distância Euclidiana, a distância de Manhattan, a distância de Canberra e a distância $\chi^2$ apresentaram bom espalhamento dos valores. Para estas distâncias, foram gerados os gráficos dos valores das distâncias obtidas (Figura 5.7), que mostraram a possível aplicabilidade destas distâncias.

Sendo assim, cada distância foi aplicada no processo de agrupamento a fim de obter a distância que apresenta melhores resultados na formação dos grupos. Os testes realizados mostraram que a distância de Canberra e a distância $\chi^2$ não são aplicáveis, pois apresentaram alta variabilidade intra-grupos, ou seja, os grupos formados foram compostos por quadros não-similares. Entretanto, a distância Euclidiana e a distância de Manhattan apresentaram resultados muito bons, sendo que a distância Euclidiana foi superior a distância de Manhattan. Logo, os experimentos finais foram executados utilizando a distância Euclidiana.

### Avaliação dos Resumos através da CRU

Para avaliar os resumos produzidos pela abordagem VSUMM, foi aplicado o método CRU (Comparação com os Resumos dos Usuários). A criação dos resumos dos 50 vídeos listados na Tabela 5.2, foi realizada por 50 usuários, sendo que cada usuário gerou cinco resumos, ou seja, para cada vídeo foram criados pelos usuários cinco resumos diferentes. Dessa forma, foram obtidas 250 amostras de resumos de vídeos.

Duas abordagens são aplicadas para produzir os resumos avaliados nestes experimentos: VSUMM1 e VSUMM2. As duas apresentam os mesmos passos da metodologia descrita (VSUMM), sendo que a única diferença entre estas é a forma da identificação dos quadros-chave. Em VSUMM1, um quadro-chave é selecionado em cada grupo, e em VSUMM2, um quadro-chave é identificado em cada grupo-chave. Estas abordagens foram comparadas com outras duas abordagens da literatura – DT (Mundur et al., 2006) e VISTO (Furini et al., 2007) –, as quais aplicaram os seus métodos de sumarização nos mesmos vídeos utilizados nestes experimentos. Além destas abordagens, os resumos produzidos por VSUMM1 e VSUMM2 foram comparados com os resumos gerados pelo OV. Todos os resumos são mostrados no Apêndice A.

A qualidade dos resumos foi medida através da taxa de acerto $\text{CRU}_{acerto}$ (Equação 4.2) e da taxa de erro $\text{CRU}_{erro}$ (Equação 4.3). Nas Tabelas 5.13 e 5.14 são apresentadas as médias das taxas de acerto e de erro obtidas, respectivamente, por cada abordagem avaliada nestes experimentos (DT, OV, VISTO, VSUMM1 e VSUMM2). Nas Figuras 5.8–5.12, as taxas de acerto e de erro são graficamente representadas.

De acordo com os resultados mostrados nas tabelas supracitadas, a abordagem VSUMM1 apresenta a maior taxa de acerto em relação às outras abordagens e a abordagem VSUMM2 apresenta a menor taxa de erros. Para validar estes resultados, foi feita a comparação entre as taxas de acerto e as taxas de erro alcançadas pela abordagem VSUMM1 e as demais abordagens. Nas Tabelas 5.16 e 5.15 são apresentadas estas comparações, respectivamente.

Tabela 5.13: Média da taxa de acerto $CRU_{acerto}$ calculada para diferentes abordagens.

| **Vídeos** | **DT** | **OV** | **VISTO** | **VSUMM1** | **VSUMM2** |
|---|---|---|---|---|---|
| *v21* | 0,20 | 0,55 | **0,83** | 0,79 | 0,75 |
| *v22* | 0,17 | 0,83 | **1,00** | 0,56 | 0,56 |
| *v23* | 0,57 | 0,78 | 0,52 | **0,98** | 0,93 |
| *v24* | 0,18 | 0,75 | 0,57 | **0,89** | 0,72 |
| *v25* | 0,15 | 0,33 | 0,38 | **0,66** | 0,61 |
| *v26* | 0,38 | 0,13 | 0,66 | **0,71** | 0,53 |
| *v27* | 0,48 | 0,25 | 0,44 | **0,78** | 0,60 |
| *v28* | 0,24 | 0,63 | 0,43 | **0,82** | 0,68 |
| *v29* | 0,56 | 0,48 | 0,54 | **0,75** | **0,75** |
| *v30* | 0,58 | 0,62 | 0,60 | **0,75** | **0,75** |
| *v31* | 0,53 | 0,64 | 0,60 | **0,84** | 0,63 |
| *v32* | 0,18 | 0,53 | 0,62 | **0,81** | 0,77 |
| *v33* | 0,22 | 0,56 | 0,55 | **0,76** | 0,51 |
| *v34* | 0,60 | **1,00** | 0,94 | 0,90 | 0,70 |
| *v35* | 0,30 | 0,80 | 0,42 | **0,87** | 0,69 |
| *v36* | 0,29 | **0,89** | 0,85 | 0,77 | 0,63 |
| *v37* | 0,60 | 0,60 | **1,00** | **1,00** | 0,60 |
| *v38* | 0,38 | 0,60 | 0,79 | **0,84** | 0,55 |
| *v39* | 0,44 | 0,77 | **0,93** | 0,83 | 0,46 |
| *v40* | 0,48 | **0,98** | 0,66 | 0,82 | 0,69 |
| *v41* | 0,47 | **0,98** | 0,62 | 0,89 | 0,71 |
| *v42* | 0,54 | **0,93** | 0,66 | 0,74 | 0,42 |
| *v43* | 0,59 | **1,00** | 0,88 | 0,93 | 0,63 |
| *v44* | 0,45 | 0,91 | 0,76 | **0,93** | 0,76 |
| *v45* | 0,25 | 0,82 | 0,78 | **0,84** | 0,73 |
| *v46* | 0,66 | 0,54 | 0,51 | **0,92** | 0,86 |
| *v47* | 0,96 | **1,00** | **1,00** | 0,96 | 0,76 |
| *v48* | 0,92 | **1,00** | **1,00** | 0,92 | 0,92 |
| *v49* | 0,61 | **1,00** | 0,58 | 0,93 | 0,67 |
| *v50* | 0,65 | **1,00** | **1,00** | 0,89 | 0,85 |
| *v51* | 0,72 | 0,85 | 0,74 | **0,88** | 0,50 |
| *v52* | 0,41 | **1,00** | **1,00** | 0,95 | 0,75 |
| *v53* | 0,55 | 0,03 | **0,67** | 0,50 | 0,42 |
| *v54* | 0,51 | 0,74 | **0,89** | 0,85 | 0,85 |
| *v55* | 0,36 | 0,23 | 0,61 | **0,75** | **0,75** |
| *v56* | 0,45 | 0,69 | 0,61 | **0,96** | 0,69 |
| *v57* | 0,59 | 0,73 | 0,88 | **0,90** | 0,86 |
| *v58* | 0,65 | 0,44 | 0,74 | **0,90** | 0,75 |
| *v59* | 0,69 | 0,72 | 0,70 | **1,00** | 0,81 |
| *v60* | 0,64 | 0,83 | 0,55 | **0,93** | 0,81 |
| *v61* | 0,81 | **1,00** | 0,89 | 0,93 | 0,74 |
| *v62* | **1,00** | 0,78 | **1,00** | **1,00** | 0,78 |
| *v63* | 0,64 | 0,59 | 0,36 | **0,89** | 0,70 |
| *v64* | 0,51 | 0,65 | 0,72 | **0,76** | 0,67 |
| *v65* | 0,61 | 0,58 | 0,57 | **0,98** | 0,91 |
| *v66* | 0,77 | 0,61 | 0,68 | **0,87** | **0,87** |
| *v67* | **0,93** | 0,69 | 0,79 | **0,93** | 0,80 |
| *v68* | 0,69 | 0,65 | 0,77 | **0,88** | 0,65 |
| *v69* | 0,49 | 0,54 | 0,67 | **0,94** | 0,73 |
| *v70* | **0,96** | **0,96** | 0,84 | 0,75 | 0,75 |
| **Média** | 0,53 | 0,70 | 0,72 | **0,85** | 0,70 |

Tabela 5.14: Média da taxa de erro $CRU_{erro}$ calculada para diferentes abordagens.

| Vídeos | DT | OV | VISTO | VSUMM1 | VSUMM2 |
|---|---|---|---|---|---|
| *v21* | **0,37** | 0,73 | 1,02 | 1,35 | 0,96 |
| *v22* | 0,82 | 0,40 | 0,73 | **0,43** | **0,43** |
| *v23* | **0,30** | 0,58 | 0,35 | 0,88 | 0,68 |
| *v24* | 0,50 | 0,39 | 0,57 | 0,36 | **0,30** |
| *v25* | 0,18 | **0,09** | 0,12 | 0,17 | 0,14 |
| *v26* | **0,08** | 0,14 | 0,65 | 0,40 | 0,25 |
| *v27* | **0,45** | 0,55 | 0,89 | 0,94 | 0,59 |
| *v28* | **0,08** | 0,12 | 0,15 | 0,19 | 0,17 |
| *v29* | 0,55 | **0,31** | 0,41 | 0,68 | 0,52 |
| *v30* | 0,16 | 0,11 | 0,13 | **0,09** | **0,09** |
| *v31* | **0,16** | **0,16** | 0,32 | 0,42 | 0,28 |
| *v32* | **0,19** | 0,45 | 0,60 | 0,54 | 0,58 |
| *v33* | **0,20** | 0,36 | 0,59 | 0,67 | 0,41 |
| *v34* | 0,10 | 0,70 | 0,46 | **0,00** | **0,00** |
| *v35* | 0,31 | 0,73 | 0,29 | 0,26 | **0,22** |
| *v36* | **0,83** | 1,17 | 1,76 | 1,28 | 0,87 |
| *v37* | 0,20 | 0,20 | 1,20 | **0,00** | **0,00** |
| *v38* | 0,34 | 0,30 | 0,57 | 0,43 | **0,26** |
| *v39* | **0,05** | 0,30 | 0,46 | 0,09 | 0,13 |
| *v40* | **0,32** | 1,39 | 0,53 | 0,89 | 0,63 |
| *v41* | 0,68 | 0,81 | 0,53 | 0,39 | **0,19** |
| *v42* | **0,08** | 0,81 | 0,33 | 0,13 | **0,08** |
| *v43* | 0,17 | 1,11 | 0,33 | 0,28 | **0,13** |
| *v44* | **0,11** | 0,22 | 0,25 | 0,30 | 0,14 |
| *v45* | **0,24** | 1,14 | 0,52 | 0,30 | 0,25 |
| *v46* | 0,44 | 0,42 | 1,00 | 0,46 | **0,37** |
| *v47* | **0,04** | 0,40 | 0,40 | **0,04** | **0,04** |
| *v48* | **0,12** | 0,93 | 0,79 | **0,12** | **0,12** |
| *v49* | 0,17 | 0,71 | 0,35 | 0,08 | **0,03** |
| *v50* | 0,22 | 1,18 | 1,33 | 0,27 | **0,17** |
| *v51* | 0,40 | 1,25 | 0,52 | 0,24 | **0,07** |
| *v52* | 0,10 | 0,76 | 0,51 | 0,05 | **0,00** |
| *v53* | 0,70 | 3,47 | 0,83 | 0,75 | **0,58** |
| *v54* | 0,40 | 0,46 | 0,47 | **0,21** | **0,21** |
| *v55* | 0,53 | **0,36** | 1,17 | 0,43 | 0,43 |
| *v56* | 0,44 | 0,38 | 0,64 | 0,48 | **0,20** |
| *v57* | 0,45 | **0,31** | 0,76 | 0,44 | 0,34 |
| *v58* | 0,21 | **0,09** | 0,33 | 0,49 | 0,22 |
| *v59* | 0,16 | **0,03** | 0,16 | 0,28 | 0,05 |
| *v60* | 0,36 | 0,41 | 0,32 | 0,44 | **0,31** |
| *v61* | 0,29 | 1,62 | 0,86 | 0,60 | **0,13** |
| *v62* | 0,13 | 1,20 | 0,98 | 0,13 | **0,07** |
| *v63* | 0,20 | 0,25 | 0,61 | 0,36 | **0,14** |
| *v64* | 0,29 | **0,22** | 0,51 | 0,33 | 0,28 |
| *v65* | 0,46 | **0,22** | 0,63 | 0,49 | 0,43 |
| *v66* | 0,30 | 0,19 | 0,25 | 0,33 | 0,33 |
| *v67* | **0,00** | 0,11 | 0,41 | **0,00** | **0,00** |
| *v68* | 0,23 | 0,27 | 0,61 | **0,04** | **0,04** |
| *v69* | 0,13 | **0,07** | 0,78 | 0,15 | 0,12 |
| *v70* | **0,09** | 0,09 | 0,21 | 0,30 | 0,30 |
| **Média** | 0,29 | 0,57 | 0,58 | 0,38 | **0,27** |

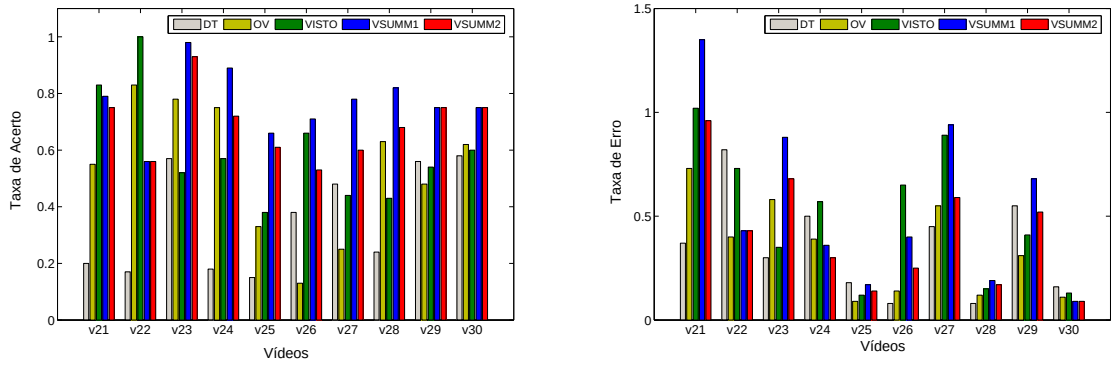



Figura 5.8: Gráfico das taxas de acerto e de erro para os vídeos *v21* a *v30*.

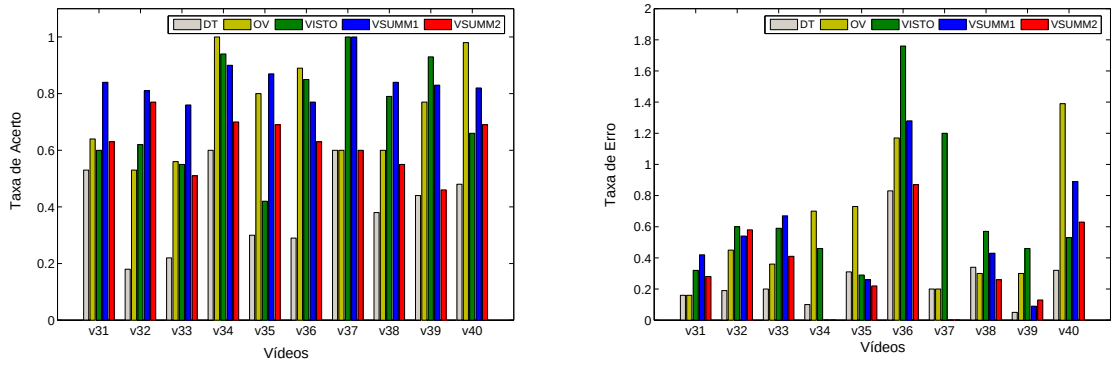



Figura 5.9: Gráfico das taxas de acerto e de erro para os vídeos *v31* a *v40*.

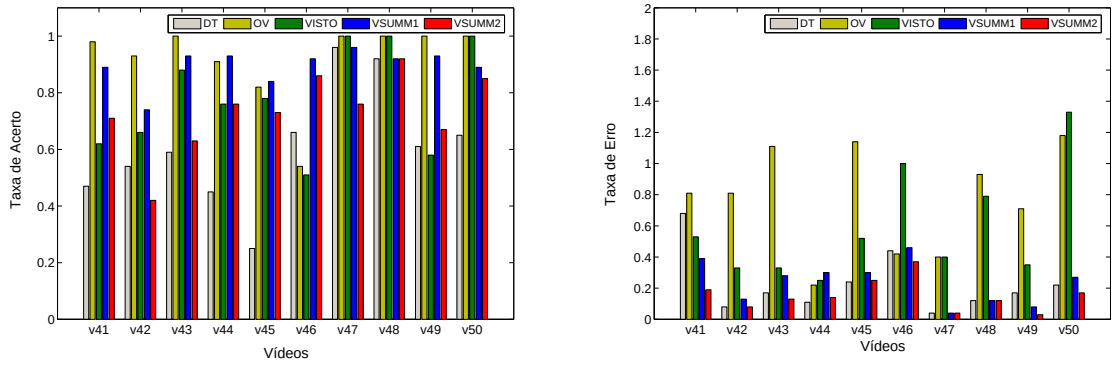



Figura 5.10: Gráfico das taxas de acerto e de erro para os vídeos *v41* a *v50*.

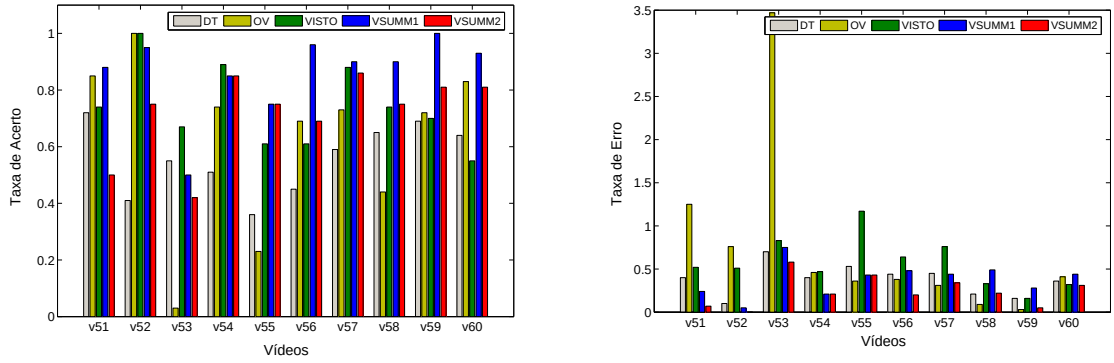



Figura 5.11: Gráfico das taxas de acerto e de erro para os vídeos *v51* a *v60*.

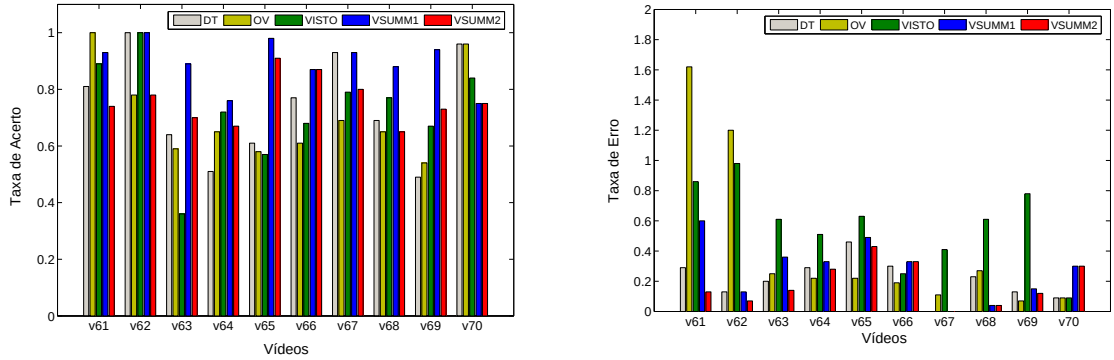



Figura 5.12: Gráfico das taxas de acerto e de erro para os vídeos *v61* a *v70*.

Tabela 5.15: Comparação entre as taxas de erro $\text{CRU}_{erro}$ calculadas para as diferentes abordagens.

| **Diferença** | **Média** | **Desvio Padrão** | **Intervalo de Confiança (98%)** | |
|---|---|---|---|---|
| | | | **mínimo** | **máximo** |
| VSUMM1 – DT | 0,09 | 0,24 | 0,01 | 0,17 |
| VSUMM1 – OV | –0,19 | 0,56 | –0,38 | –0,01 |
| VSUMM1 – VISTO | –0,20 | 0,34 | –0,32 | –0,09 |
| VSUMM1 – VSUMM2 | 0,11 | 0,12 | 0,07 | 0,15 |

Tabela 5.16: Comparação entre as taxas de acerto $\text{CRU}_{acerto}$ calculadas para as diferentes abordagens.

| **Diferença** | **Média** | **Desvio Padrão** | **Intervalo de Confiança (98%)** | |
|---|---|---|---|---|
| | | | **mínimo** | **máximo** |
| VSUMM1 – DT | 0,32 | 0,19 | 0,26 | 0,38 |
| VSUMM1 – OV | 0,15 | 0,21 | 0,08 | 0,22 |
| VSUMM1 – VISTO | 0,14 | 0,19 | 0,07 | 0,20 |
| VSUMM1 – VSUMM2 | 0,15 | 0,11 | 0,11 | 0,18 |

Com 98% de confiança, pode-se afirmar que a abordagem VSUMM1 fornece resultados com qualidade superior (maior taxa de acerto) em relação às demais abordagens avaliadas. Ou ainda, os resumos produzidos pela abordagem VSUMM1 são os mais próximos dos resumos criados pelos usuários. Além disso, para este mesmo nível de confiança, também pode-se afirmar que a abordagem VSUMM1 apresenta uma taxa de erro menor em relação às abordagens OV e VISTO, e apresenta, no entanto, uma taxa de erro maior em relação às abordagens DT e VSUMM2.

No caso da abordagem DT, este resultado enganosamente favorável era previsto devido aos resumos produzidos por esta abordagem. O tamanho dos resumos DT é muito menor em relação ao tamanho dos resumos criados pelos usuários, conseqüentemente os resumos DT apresentam uma baixa taxa de erro ao custo de uma baixa taxa de acerto. Logo, apesar da taxa de erro da abordagem DT ser estatisticamente menor em relação à taxa de erro da abordagem VSUMM1, este resultado pode ser desconsiderado, pois não é interessante obter resumos que apresentam baixa taxa de erro e ao mesmo tempo baixa taxa de acerto.

No caso da abordagem VSUMM2, a análise é similar à abordagem DT. Os resumos produzidos através abordagem VSUMM2 são no máximo do mesmo tamanho dos resumos gerados pela abordagem VSUMM1. Como VSUMM2 produz resumos menores, tende a errar menos, mas também a acertar menos, como pode ser visto na Tabela 5.13, onde a taxa de acerto alcançada pelos resumos gerados através da abordagem VSUMM2 é consideravelmente menor que a taxa obtida pelos resumos produzidos através da abordagem VSUMM1.

Diante destas observações, pode-se concluir que a abordagem VSUMM1 fornece resultados melhores em relação às abordagens avaliadas neste trabalho.

### Análise dos Resultados Finais

Conforme os resultados experimentais obtidos, observou-se que a abordagem proposta apresentou resumos com qualidade superior às demais abordagens encontradas na literatura. Um ponto importante verificado foi as taxas de acerto obtidas pelas abordagens VSUMM1 e VSUMM2. Ao contrário do que era esperado, a abordagem VSUMM1 forneceu resultados com qualidade superior em relação à abordagem VSUMM2. Como esta abordagem seleciona os quadros-chave apenas dos grupos-chave, eliminando os grupos que em teoria não teriam tanta importância, pois são compostos por um número relativamente pequeno de quadros, então esperava-se que a taxa de acerto da abordagem VSUMM2 fosse maior que a taxa de acerto da abordagem VSUMM1.

Outro fato observado, que implicou na verificação anterior, foi o tamanho razoavelmente grande dos resumos criados pelos usuários. Antes de realizar o processo de avaliação, foi informado aos usuários que eles deveriam selecionar os quadros que, na opinião deles, conseguiriam representar o conteúdo do vídeo de forma concisa. Desta maneira, esperava-se obter resumos que continham os quadros-chave mais relevantes do vídeo. No entanto, os resumos criados pelos usuários mostraram a necessidade do usuário em obter resumos mais extensos, que representem vários segmentos do vídeo, independentemente do tamanho destes segmentos.

## 5.4 Considerações

Durante a elaboração deste trabalho, diversos experimentos foram realizados. Os experimentos preliminares foram importantes para estudar o comportamento da metodologia proposta e principalmente para identificar os problemas existentes e apontar soluções. A partir dos problemas identificados, algumas soluções foram propostas e então novos experimentos foram realizados. Os resultados obtidos validaram o método proposto para a sumarização automática de vídeos.

# Capítulo 6

# Conclusões e Trabalhos Futuros

A sumarização automática de vídeos é uma área que vem recebendo crescente atenção por parte da comunidade científica. Esse interesse pode ser explicado por vários fatores como, por exemplo, (1) a redução do custo de equipamentos de captura, transmissão e armazenamento de vídeos; (2) o crescimento exponencial do número de vídeos publicados na Internet; (3) os desafios científicos envolvidos, (4) as inúmeras aplicações práticas em sistemas como máquinas de busca, bibliotecas digitais e sistemas de segurança e (5) a inadequação de técnicas tradicionais de sumarização de vídeo para descrever, representar e realizar buscas em grandes coleções de vídeos como, por exemplo, os serviços de busca de vídeos do Alta Vista[1], do Yahoo[2] e da Google[3] que representam todo o vídeo com apenas um quadro.

Neste trabalho, foi apresentada uma metodologia para a elaboração automática de resumos estáticos de vídeos. A metodologia proposta integrou as vantagens dos conceitos apresentados nos trabalhos de referência na área de sumarização de vídeos de modo a aproveitar em um único método as contribuições relevantes de técnicas anteriormente propostas, contribuindo para o aprimoramento do estado da arte desta área de aplicação.

Mais especificamente, este trabalho se concentrou na classificação não-supervisionada dos quadros dos vídeos baseada na informação de cor presente nestes. A cor é uma das características visuais mais amplamente utilizadas na descrição de imagens por ser simples, intuitiva, estar presente na maioria das imagens e fornecer bons resultados. A descrição do conteúdo visual do vídeo foi feita utilizando histogramas de cor e a classificação dos quadros foi realizada por meio do algoritmo de agrupamento *k-means*.

A solução proposta apresentou, com um nível de confiança de 98%, resultados com qualidade superior em relação às demais abordagens encontradas na literatura.

## 6.1 Objetivos Alcançados

Neste trabalho, foi definida uma metodologia capaz de produzir resumos estáticos de vídeos de forma eficaz e eficiente. Para isso, foram avaliadas técnicas simples para identificar as

[1] http://video.altavista.com
[2] http://video.search.yahoo.com
[3] http://video.google.com

características visuais de cor apropriadas para o processo de sumarização proposto.

Um amplo levantamento bibliográfico, com o estudo dos principais métodos que compõem o estado da arte no que se refere à elaboração de resumos estáticos de vídeos, foi realizado, o qual contribuiu na definição da solução proposta.

Também, foi definido um método de avaliação de resumos que permitiu quantificar a qualidade dos resumos gerados, reduzir a subjetividade presente na tarefa de avaliação e principalmente, realizar de forma rápida comparações entre resumos de diferentes abordagens.

## 6.2 Trabalhos Futuros

O presente trabalho pode ser continuado explorando-se os seguintes aspectos:

- Verificar a adequação da metodologia proposta a outros gêneros de vídeos, tais como caseiro, desenho animado, esportivo, filme; e testar o funcionamento do método em vídeos com tempo de duração maior.

- Empregar outras características visuais, como por exemplo, textura, forma, movimento e relações topológicas entre regiões do quadro. A combinação destas características pode levar a melhores resultados. Combinar características oriundas de domínios diferentes, como histogramas de cor (domínio de cor) e variação entre *pixels* (domínio espacial) pode ser uma forma de compensar suas deficiências individuais.

- Utilizar outras abordagens para extrair as características visuais dos quadros. As abordagens existentes para este processo podem ser classificadas em (Stehling et al., 2002): (1) globais (neste trabalho foi aplicada uma abordagem global), (2) baseadas em particionamento e (3) regionais.

- Investigar o uso de outras técnicas de classificação, como por exemplo o algoritmo DBSCAN (Ester et al., 1996), que utiliza a idéia de densidade de objetos para selecionar as áreas com maior densidade de objetos, podendo-se obter áreas com formatos variados e não somente circulares como é o caso do *k-means*.

- Empregar outros métodos para determinar o número de ideal de grupos para cada vídeo. Existem diversos trabalhos nesta direção, tais como (Ishioka, 2000; Pelleg e Moore, 2000; Guo et al., 2006; Chiang e Mirkin, 2007), mas o problema continua em aberto. Algumas das técnicas utilizadas para determinar o número de grupos são o critério de informação de Akaike (do inglês, *Akaike's Information Criterion* – AIC) (Akaike, 1974) e critério do tamanho mínimo de descrição (do inglês, *Minimum Description Length* – MDL) (Rissanen, 1978), que também podem ser aplicados neste trabalho.

- Estender a metodologia proposta para gerar resumos dinâmicos. Esses podem ser criados a partir de quadros-chave por meio da junção de segmentos de tamanho fixo, subtomadas ou da tomada à qual o quadro-chave pertence (Wu et al., 2000; Lee e Hayes, 2004).

## 6.3 Publicações oriundas deste Trabalho

- Avila, S. E. F.; Luz Jr, A.; Araújo (2008) VSUMM: A Simple and Efficient Approach for Automatic Video Summarization. 15th International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 449–452, Bratislava, Slovak Republic. http://dx.doi.org/10.1109/IWSSIP.2008.4604463. Qualis A International.

- Avila, S. E. F.; Luz Jr, A.; Araújo, A. A.; Cord, M. (2008) VSUMM: An Approach for Automatic Video Summarization and Quantitative Evaluation. 21th Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI), Campo Grande, Brazil, pp. 103–110, IEEE Computer Society. http://dx.doi.org/10.1109/SIBGRAPI.2008.31. Qualis A Nacional.

# Referências Bibliográficas


Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723. 66

Androutsos, D.; Plataniotiss, K. N. e Venetsanopoulos, A. N. (1998). Distance measures for color image retrieval. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 2, pp. 770–774. http://dx.doi.org/10.1109/ICIP.1998.723652. 17

Avrithis, Y. S.; Doulamis, A. D.; Doulamis, N. D. e Kollias, S. D. (1999). A stochastic framework for optimal key frame extraction from `MPEG` video databases. *Computer Vision and Image Understanding*, 75(1-2):3–24. http://dx.doi.org/10.1006/CVIU.1999.0761. 27

Beaver, F. E. (1994). *Dictionary of Film Terms*. Twayne Publishing NY. 8

Boreczky, J. S. e Rowe, L. A. (1996). Comparison of video shot boundary detection techniques. In *Proceedings of the IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases*, volume SPIE 2670, pp. 170–179. http://dx.doi.org/10.1117/12.238675. 21

Ćalić, J.; Gibson, D. P. e Campbell, N. W. (2007). Efficient layout of comic-like video summaries. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(7):931–936. http://dx.doi.org/10.1109/TCSVT.2007.897466. 2

Ćalić, J. e Izquierdo, E. (2002). Efficient key-frame extraction and video analysis. In *Proceedings of the International Conference on Information Technology*, p. 28, Washington, DC, USA. IEEE Computer Society. http://dx.doi.org/10.1109/ITCC.2002.1000355. 28

Cámara-Chávez, G. (2007). *Análise de Conteúdo de Vídeo por Meio do Aprendizado Ativo*. PhD thesis, Universidade Federal de Minas Gerais. 17, 37, 55

Cámara-Chávez, G.; Precioso, F.; Cord, M.; Phillip-Foliguet, S. e de A. Araújo, A. (2007). Shot boundary detection by a hierarchical supervised approach. *International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 197–200. http://dx.doi.org/10.1109/IWSSIP.2007.4381187. 21

——— (2008). An interactive video content-based retrieval system. *International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 133–136. http://dx.doi.org/10.1109/IWSSIP.2008.4604385. 20

Chang, H. S.; Sulla, S. H. e Lee, S. U. (1999). Efficient video indexing scheme for content-based retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(8):1269–1279. http://dx.doi.org/10.1109/76.809161. 24, 31

Chang, I.-C. e Chen, K.-Y. (2007). Content-selection based video summarization. *Digest of Technical Papers International Conference on Consumer Electronics (ICCE)*, pp. 1–2. http://dx.doi.org/10.1109/ICCE.2007.341528. 2, 20, 21

Chiang, M. M.-T. e Mirkin, B. (2007). Experiments for the number of clusters in K-means. In *Progress in Artificial Intelligence, 13th Portuguese Conference on Artificial Intelligence (EPIA)*, volume 4874 of *Lecture Notes in Computer Science*, pp. 395–405. Springer. http://dx.doi.org/10.1007/978-3-540-77002-2 33. 66

Ciocca, G. e Schettini, R. (2005). Dynamic key-frame extraction for video summarization. In *Proceedings of the SPIE International Society for Optical Engineering*, volume 5670, pp. 137–142. 20

——— (2006). Visual video summaries post-processing using supervised and unsupervised classification. *Digest of Technical Papers IEEE International Conference on Consumer Electronics (ICCE)*, pp. 81–82. http://dx.doi.org/10.1109/ICCE.2006.1598320. 25

Cooper, M. e Foote, J. (2005). Discriminative techniques for keyframe selection. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 502–505. http://dx.doi.org/10.1109/ICME.2005.1521470. 25

Cotsaces, C.; Nikolaidis, N. e Pitas, L. (2006). Video shot detection and condensed representation: A review. *IEEE Signal Processing Magazine*, 23(2):28–37. http://dx.doi.org/10.1109/MSP.2006.1621446. 21

da Silva Torres, R. e Falcão, A. X. (2006). Content-based image retrieval: Theory and applications. *Revista de Informática Teórica e Aplicada (RITA)*, 13(2):161–185. 10

DeMenthon, D.; Kobla, V. e Doermann, D. (1998). Video summarization by curve simplification. In *Proceedings of the ACM International Conference on Multimedia*, pp. 211–218, New York, NY, USA. ACM. http://dx.doi.org/10.1145/290747.290773. 28, 42

Divakaran, A.; Peker, K. A. e Radhakrishnan, R. (2002). Motion activity-based extraction of key-frames from video shots. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 1, pp. 932–935. http://dx.doi.org/10.1109/ICIP.2002.1038180. 24

Doulamis, A. D.; Doulamis, N. D. e Kollias, S. D. (2000). A fuzzy video content representation for video summarization and content-based retrieval. *Signal Processing*, 80(6):1049–1067. http://dx.doi.org/10.1016/S0165-1684(00)00019-0. 27

Doulamis, N. D.; Doulamis, A. D.; Avrithis, Y. S. e Kollias, S. D. (1998). Video content representation using optimal extraction of frames and scenes. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 1, pp. 875–879. http://dx.doi.org/10.1109/ICIP.1998.723660. 27

Duda, R. O.; Hart, P. E. e Stork, D. G. (2001). *Pattern Classification*, chapter Unsupervised Learning and Clustering, p. 654. Springer-Verlag New York, Inc. 37

Dufaux, F. (2000). Key frame selection to represent a video. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 2, pp. 275–278. http://dx.doi.org/10.1109/ICIP.2000.899354. 20, 21, 29, 31

Ester, M.; Kriegel, H.-P.; Sander, J. e Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 226–231. 66

Fauvet, B.; Bouthemy, P.; Gros, P. e Spindler, F. (2004). A geometrical key-frame selection method exploiting dominant motion estimation in video. In *International Conference on Image and Video Retrieval*, Lecture Notes in Computer Science, pp. 419–427. Springer. http://dx.doi.org/10.1007/b98923. 24, 31

Foley, J. D.; van Dam, A.; Feiner, S. K. e Hughes, J. F. (1990). *Computer Graphics: Principles and Practice*. Addison-Wesley, Reading, MA, 2 edição. viii, 13, 14

Furini, M.; Geraci, F.; Montangero, M. e Pellegrini, M. (2007). VISTO: visual storyboard for web video browsing. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR)*, pp. 635–642. http://doi.acm.org/10.1145/1282280.1282370. 2, 4, 6, 10, 20, 21, 26, 31, 32, 36, 38, 43, 54, 55, 58

——— (2008). On using clustering algorithms to produce video abstracts for the web scenario. In *Proceedings of the IEEE Consumer Communication and Networking (CCNC)*, pp. 1112–1116. IEEE Communication Society. http://dx.doi.org/10.1109/ccnc08.2007.251. 2

Geraci, F.; Pellegrini, M.; Pisati, P. e Sebastiani, F. (2006). A scalable algorithm for high-quality clustering of web snippets. In *Proceedings of the ACM Symposium on Applied Computing (SAC)*, pp. 1058 –1062. http://dx.doi.org/10.1145/1141277.1141527. 6

Gibson, D.; Campbell, N. W. e Thomas, B. T. (2002). Visual abstraction of wildlife footage using gaussian mixture models and the minimum description length criterion. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, volume 814–817. http://dx.doi.org/10.1109/ICPR.2002.1048427. 25, 26

Girgensohn, A. (2003). A fast layout algorithm for visual video summaries. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*, pp. 77–80, Washington, DC, USA. IEEE Computer Society. http://dx.doi.org/10.1109/ICME.2003.1221557. 2

Girgensohn, A. e Boreczky, J. (1999). Time-constrained keyframe selection technique. *Proceedings of the IEEE International Conference on Multimedia Computing and Systems (ICMCS)*, 1:756–761. http://dx.doi.org/10.1109/MMCS.1999.779294. 25, 26

Girgensohn, A.; Boreczky, J. e Wilcox, L. (2001). Keyframe-based user interfaces for digital video. *IEEE Computer*, 34(9):61–67. http://dx.doi.org/10.1109/2.947093. 2

Gong, Y. e Liu, X. (2000). Video summarization using singular value decomposition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2174–2180, Los Alamitos, CA, USA. IEEE Computer Society. http://dx.doi.org/10.1109/CVPR.2000.854772. 2, 5, 7, 10, 20, 21, 26, 55

——— (2003). Video summarization and retrieval using singular value decomposition. *Multimedia Systems*, 9(2):157–168. http://dx.doi.org/10.1007/s00530-003-0086-3. 31

Gonzalez, R. C. e Woods, R. E. (2006). *Digital Image Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 3 edição. viii, 12, 13, 14

Guimarães, S. J. F. (2003). *Video transition identification based on 2*D *image analysis*. PhD thesis, Universidade Federal de Minas Gerais. 36

Guimarães, S. J. F.; Couprie, M.; de Albuquerque Araújo, A. e Leite, N. J. (2003). Video segmentation based on 2d image analysis. *Pattern Recognition Letters*, 24(7):947–957. 21

Guironnet, M.; Pellerin, D.; Guyader, N. e Ladret, P. (2007). Video summarization based on camera motion and a subjective evaluation method. *EURASIP Journal on Image and Video Processing*, pp. Article ID 60245, 12 pages. http://dx.doi.org/10.1155/2007/60245. 31, 32, 39, 55

Guo, H.-X.; jun Zhu, K.; wei Gao, S. e Liu, T. (2006). An improved genetic k-means algorithm for optimal clustering. In *Proceedings of the International Conference on Data Mining (ICDMW)*, pp. 793–797, Washington, DC, USA. IEEE Computer Society. http://dx.doi.org/10.1109/ICDMW.2006.30. 66

Hadi, Y.; Essannouni, F. e Thami, R. O. H. (2006). Video summarization by k-medoid clustering. In *Proceedings of the ACM Symposium on Applied Computing (SAC)*, pp. 1400–1401, New York, NY, USA. http://dx.doi.org/10.1145/1141277.1141601. 2, 20, 26, 55

Hammoud, R. I. (2006). *Interactive Video Algorithms and Technologies*. Springer Berlin Heidelberg. 35

Han, S.-H. e Kweon, I.-S. (2005). Scalable temporal interest points for abstraction and classification of video events. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 670–673. http://dx.doi.org/10.1109/ICME.2005.1521512. 30

Hanjalic, A. (2002). Shot-boundary detection: unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology*, 12(2):90–105. http://dx.doi.org/10.1109/76.988656. 20, 21

Hanjalic, A.; Lagendijk, R. L. e Biemond, J. (1998). *A new method for key frame based video content representation.* In Image Databases and Multi-Media Search. World Scientific, Singapore. 27, 30

Hanjalic, A. e Zhang, H. (1999). An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(8):1280–1289. 2, 5, 10, 20, 26, 55

Hua, X.-S.; Li, S. e Zhang, H.-J. (2005). Video booklet. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 189–192. 32

Ishioka, T. (2000). Extended K-means with an efficient estimation of the number of clusters. In *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning, Data Mining, Financial Engineering, and Intelligent Agents*, pp. 17–22, London, UK. Springer-Verlag. 66

Jain, A. K. e Dubes, R. C. (1988). *Algorithms for clustering data.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA. 16

Jain, A. K.; Murty, M. N. e Flynn, P. J. (1999). Data clustering: A review. *ACM Computer Surverys*, 31(3):264–323. http://dx.doi.org/10.1145/331499.331504. viii, 5, 15, 25

Jain, R. (1992). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling.* John Wiley and Sons, Inc. 47

Joshi, A.; Auephanwiriyakul, S. e Krishnapuram, R. (1998). On fuzzy clustering and content based access to networked video databases. In *Proceedings of the International Workshop on Research Issues in Data Engineering (RIDA) Conference*, pp. 42–43. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.56.3189. 30

Kang, E. K.; Kim, S. J. e Choi, J. S. (1999). Video retrieval based on scene change detection in compressed streams. *IEEE Transactions on Consumer Electronics*, 45(3):224–225. http://dx.doi.org/10.1109/ICCE.1999.785241. 23

Kim, C. e Hwang, J.-N. (2002). Object-based video abstraction for video surveillance systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(12):1128–1138. http://dx.doi.org/10.1109/TCSVT.2002.806813. 24

Koprinska, I. e Carrato, S. (2001). Temporal video segmentation: a survey. *Signal Processing: Image Communication*, 16(5):477–500. http://dx.doi.org/10.1117/12.333848. 20, 21

Latecki, L. J.; de Wildt, D. e Hu, J. (2001). Extraction of key frames from videos by optimal color composition matching and polygon simplification. In *Proceedings of the Multimedia Signal Processing Conference*, Cannes, France. http://dx.doi.org/10.1109/MMSP.2001.962741. 28

Lee, H. e Kim, S. (2002). Rate-driven key frame selection using temporal variation of visual content. *Electronics Letters*, 38(5):217–218. http://dx.doi.org/10.1049/el:20020112. 24

Lee, H.-C. e Kim, S.-D. (2003). Iterative key frame selection in the rate-constraint environment. *Signal Processing: Image Communication*, 18:1–15. http://dx.doi.org/10.1016/S0923-5965(02)00089-9. 27

Lee, S. e Hayes, M. H. (2003). A fast clustering algorithm for video abstraction. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 2, pp. 563–566. http://dx.doi.org/10.1109/ICIP.2003.1246742. 31

——— (2004). An application for interactive video abstraction. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pp. 905–908. http://dx.doi.org/10.1109/ICASSP.2004.1327258. 2, 66

Li, Y.; Zhang, T. e Tretter, D. (2001). An overview of video abstraction techniques. Technical report, HP Laboratory, HP-2001-191. http://serv2.ist.psu.edu:8080/viewdoc/summary?doi=10.1.1.84.6173. 1

Li, Z.; Katsaggelos, K. e Gandhi, B. (2003). Temporal rate-distortion based optimal video summary generation. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 693–696, Washington, DC, USA. http://dx.doi.org/10.1109/ICME.2003.1221711. 20, 21

Li, Z.; Schuster, G. M. e Katsaggelos, A. K. (2005). Minmax optimal video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(10):1245–1256. http://dx.doi.org/10.1109/TCSVT.2005.854230. 20

Li, Z.; Schuster, G. M.; Katsaggelos, A. K. e Gandhi, B. (2004). Optimal video summarization with a bit budget constraint. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 617–620. http://dx.doi.org/10.1109/ICIP.2004.1418830. 28

Lienhart, R. (2001). Reliable transition detection in videos: A survey and practitioner's guide. *International Journal of Image and Graphics*, 1(3):469–486. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.95.1401. 20, 21

Liu, T. e Kender, J. R. (2002a). An efficient error-minimizing algorithm for variable-rate temporal video sampling. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, volume 1, pp. 413–416. http://dx.doi.org/10.1109/ICME.2002.1035806. 27

——— (2002b). Optimization algorithms for the selection of key frame sequences of variable length. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 403–417, London, UK. Springer-Verlag. http://dx.doi.org/10.1007/3-540-47979-1 27. 27, 31

——— (2002c). Rule-based semantic summarization of instructional videos. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 1, pp. 601–604. http://dx.doi.org/10.1109/ICIP.2002.1038095. 29

Liu, T.; Zhang, H. e Qi, F. (2003). A novel video key-frame-extraction algorithm based on perceived motion energy model. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(10):1006–1013. http://dx.doi.org/10.1109/TCSVT.2003.816521. 29, 31

Liu, T.-Y.; Zhang, X.-D.; Feng, J. e Lo, K.-T. (2004). Shot reconstruction degree: a novel criterion for key frame selection. *Pattern Recognition Letters*, 25(12):1451–1457. http://dx.doi.org/10.1016/j.patrec.2004.05.020. 28

Ma, Y.-F.; Lu, L.; Zhang, H.-J. e Li, M. (2002). A user attention model for video summarization. In *Proceedings of the ACM International Conference on Multimedia*, pp. 533–542, New York, NY, USA. ACM. http://dx.doi.org/10.1145/641007.641116. 31

Ma, Y.-F. e Zhang, H.-J. (2005). Video snapshot: A bird view of video sequence. In *Proceedings of the IEEE International Multimedia Modelling Conference (MMM)*, pp. 94–101, Washington, DC, USA. IEEE Computer Society. http://dx.doi.org/10.1109/MMMC.2005.71. 20, 21

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. e Neyman, J., editores, *Proceedings of The Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297. University of California Press. 16, 18, 36

Money, A. G. e Agius, H. (2008). Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation (JVCIR)*, 19(2):121–143. http://dx.doi.org/10.1016/j.JVCIR.2007.04.002. 38

Mundur, P.; Rao, Y. e Yesha, Y. (2006). Keyframe-based video summarization using Delaunay clustering. *International Journal on Digital Libraries*, 6(2):219–232. http://dx.doi.org/10.1007/s00799-005-0129-9. 2, 4, 6, 7, 10, 20, 21, 26, 31, 32, 43, 54, 55, 58, 78

Pelleg, D. e Moore, A. W. (2000). X-means: Extending K-means with efficient estimation of the number of clusters. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 727–734, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 66

Pfeiffer, S.; Lienhart, R.; Fischer, S. e Effelsberg, W. (1996). Abstracting digital movies automatically. Technical report, University of Mannheim. 1

Porter, S. V.; Mirmehdi, M. e Thomas, B. T. (2003). A shortest path representation for video summarisation. In *Proceedings of the IEEE International Conference on Image Analysis and Processing (ICIAP)*, pp. 460–465. http://dx.doi.org/10.1109/ICIAP.2003.1234093. 20, 21, 27

Poynton, C. (2003). *Digital Video and* HDTV *Algorithms and Interfaces*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 8

Rao, Y.; Mundur, P. e Yesha, Y. (2004). Automatic video summarization for wireless and mobile environments. *IEEE International Conference on Communications (ICC)*, 3:1532–1536. http://dx.doi.org/10.1109/ICC.2004.1312767. 21

Rasheed, Z. e Shah, M. (2003). Scene detection in hollywood movies and tv shows. In *International Conference on Computer Vision and Pattern Recognition*, pp. 343–350. IEEE Computer Society. http://dx.doi.org/10.1109/CVPR.2003.1211489. 24

Rissanen, J. (1978). Modelling by shortest data description. *Automatica*, 14:465–471. 66

Rong, J.; Jin, W. e Wu, L. (2004). Key frame extraction using inter-shot information. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 571–574. http://dx.doi.org/10.1109/ICME.2004.1394256. 20, 25

Stehling, R. O.; Nascimento, M. A. e ao, A. X. F. (2002). *Multimedia Mining: a High Way to Intelligent Multimedia Document*, chapter Techniques for color-based image retrieval, pp. 61–80. Kluwer Academic. 35, 66

Sun, X. e Kankanhalli, M. S. (2000). Video summarization using R-sequences. *Real-Time Imaging*, 6(6):449–459. http://dx.doi.org/10.1006/RTIM.1999.0197. 20, 21, 24

Swain, M. J. e Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, 7(1):11–32. http://dx.doi.org/10.1007/BF00130487. 10, 17

Tonomura, Y.; Akutsu, A.; Otsuji, K. e Sadakata, T. (1993). VideoMAP and VideoSpaceIcon: Tools for anatomizing video content. In *Proceedings of the INTERCHI Conference on Human Factors in Computing Systems*, pp. 131–136, Amsterdam, The Netherlands. IOS Press. http://dx.doi.org/10.1145/169059.169117. 2

Truong, B. T. e Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3(1). http://dx.doi.org/10.1145/1198302.1198305. viii, 1, 23, 26, 28, 29, 30

Uchihashi, S.; Foote, J.; Girgensohn, A. e Boreczky, J. S. (1999). Video manga: generating semantically meaningful video summaries. In *Proceedings of the ACM International Conference on Multimedia (Part 1)*, pp. 383–392, New York, NY, USA. http://dx.doi.org/10.1145/319463.319654. 2, 20, 21, 26, 55

Vermaak, J.; Pérez, P.; Gangnet, M. e Blake, A. (2002). Rapid summarisation and browsing of video sequences. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 1. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.7883. 30

Wang, T.; Mei, T.; Hua, X.-S.; Liu, X.-L. e Zhou, H.-Q. (2007). Video collage: A novel presentation of video sequence. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1479–1482. http://dx.doi.org/10.1109/ICME.2007.4284941. 2, 31, 32, 38

Woods, J. W. (2006). *Multidimensional Signal, Image, and Video Processing and Coding*. Academic Press, Inc., Orlando, FL, USA. 8

Wu, J.-K. K.; Hong, D.; Kankanhalli, M. S. e Lim, J.-H. (2000). *Perspectives on Content-Based Multimedia Systems*. Kluwer Academic Publishers, Norwell, MA, USA. 2, 66

Xiong, W.; Lee, J. C.-M. e Ma, R.-H. (1997). Automatic video data structuring through shot partitioning and key-frame computing. *Machine Vision and Applications*, 10(2):51–65. http://dx.doi.org/10.1007/s001380050059. 25

Yahiaoui, I.; Mérialdo, B. e Huet, B. (2001). Automatic video summarization. In *Multimedia Content-Based Indexing and Retrieval (MCBIR)*. 20, 21, 24, 31, 32

Yeung, M. M. e Leo, B.-L. (1997). Video visualization for compact representation and fast browsing of pictorial content. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(5):771–785. 2, 32

Yeung, M. M. e Liu, B. (1995). Efficient matching and clustering of video shots. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 1, pp. 338–341. http://dx.doi.org/10.1109/ICIP.1995.529715. 20, 21, 23

Yu, X.-D.; Wang, L.; Tian, Q. e Xue, P. (2004). Multi-level video representation with application to keyframe extraction. In *Proceedings of the International Multimedia Modelling Conference*, pp. 117–121. http://dx.doi.org/10.1109/MULMM.2004.1264975. 25, 26, 30, 55

Yuan, J.; Wang, H.; Xiao, L.; Zheng, W.; Li, J.; Lin, F. e Zhang, B. (2007). A formal study of shot boundary detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(2):168–186. http://dx.doi.org/10.1109/TCSVT.2006.888023. 21

Zhang, H. J.; Wu, J.; Zhong, D. e Smoliar, S. W. (1997). An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30(4):643–658. http://dx.doi.org/10.1016/S0031-3203(96)00109-4. 23

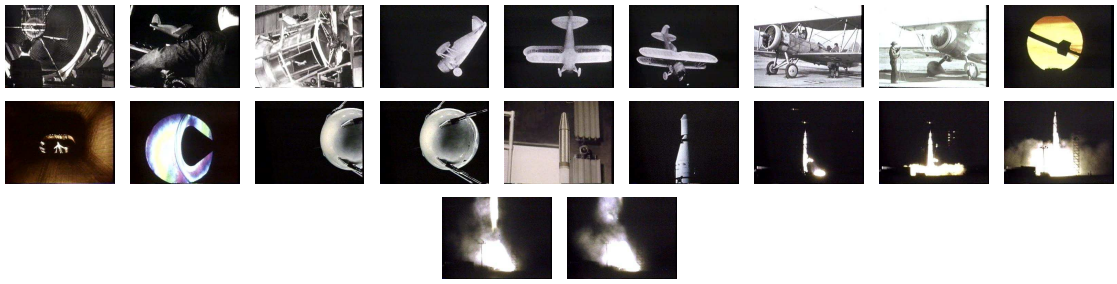
Zhang, X.-D.; Liu, T.-Y.; Lo, K.-T. e Feng, J. (2003). Dynamic selection and effective compression of key frames for video abstraction. *Pattern Recognition Letters*, 24(9–10):1523–1532. http://dx.doi.org/10.1016/S0167-8655(02)00391-4. 24, 30

Zhuang, Y.; Rui, Y.; Huang, T. S. e Mehrotra, S. (1998). Adaptive key frame extraction using unsupervised clustering. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 1, pp. 866–870. http://dx.doi.org/10.1109/ICIP.1998.723655. 2, 5, 10, 20, 25, 26, 30, 37, 54, 55

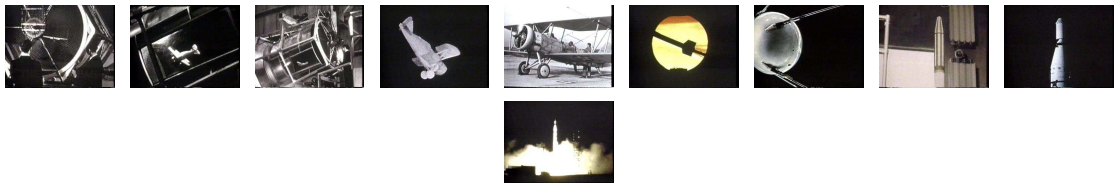
# Apêndice A

# Resumos Produzidos

Neste apêndice, são mostrados os resumos produzidos pelas diferentes abordagens. Nas Figuras A.1–A.20, são dispostos os resumos gerados pelo Open Video e pela metodologia proposta inicialmente, como também o valor da PMO obtido por cada abordagem em cada resumo. Nas Figuras A.21–A.70, são apresentados os resumos produzidos pelo Open Video e pelas abordagens DT (Mundur et al., 2006), VISTO (Furini et al., 2007), VSUMM1 e VSUMM2; e as taxas de acerto ($\text{CRU}_{acerto}$) e erro ($\text{CRU}_{erro}$) alcançadas por cada técnica em cada resumo.

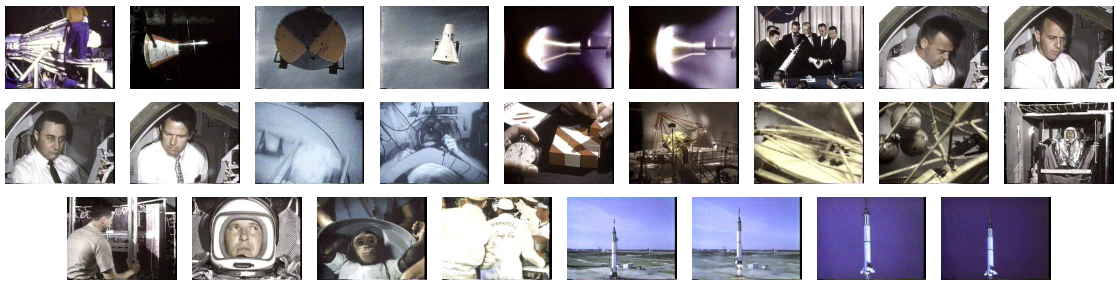


(a) Open Video. PMO = 4,0.



(b) Método Proposto. PMO = 4,4.

Figura A.1: Resumos gerados pelas diferentes abordagens para o vídeo *NASA 25th Anniversary Show, Segment 02 (v01)*.



(a) Open Video. PMO = 3,6.

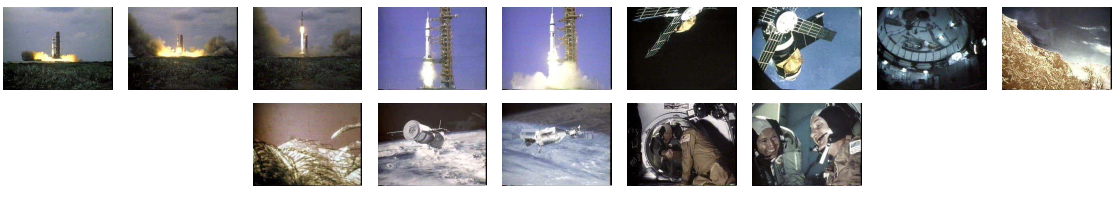
(b) Método Proposto. PMO = 3,8.

Figura A.2: Resumos gerados pelas diferentes abordagens para o vídeo *NASA 25th Anniversary Show, Segment 03 (v02)*.

(a) Open Video. PMO = 3,8.

(b) Método Proposto. PMO = 3,8

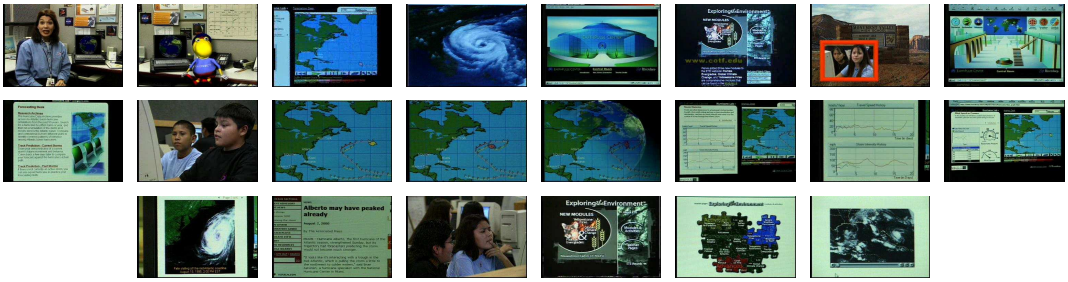
Figura A.3: Resumos gerados pelas diferentes abordagens para o vídeo *NASA 25th Anniversary Show, Segment 04 (v03)*.



(a) Open Video. PMO = 3,8.

(b) Método Proposto. PMO = 3,8.

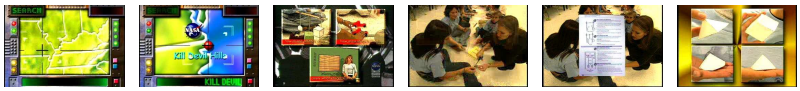
Figura A.4: Resumos gerados pelas diferentes abordagens para o vídeo *NASA 25th Anniversary Show, Segment 08 (v04)*.



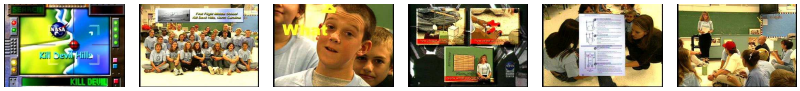
(a) Open Video. PMO = 3,8.

(b) Método Proposto. PMO = 3,8.

Figura A.5: Resumos gerados pelas diferentes abordagens para o vídeo *Hurricanes and Computer Simulation (v05)*.



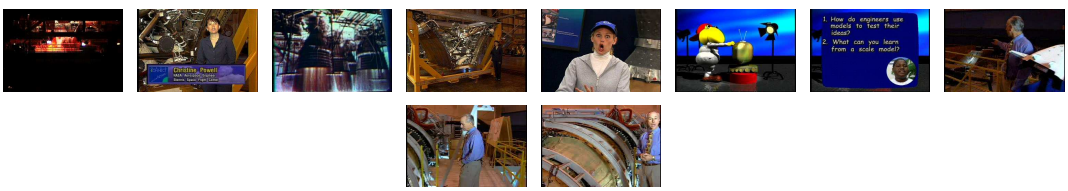
(a) Open Video. PMO = 3,4.



(b) Método Proposto. PMO = 3,3.

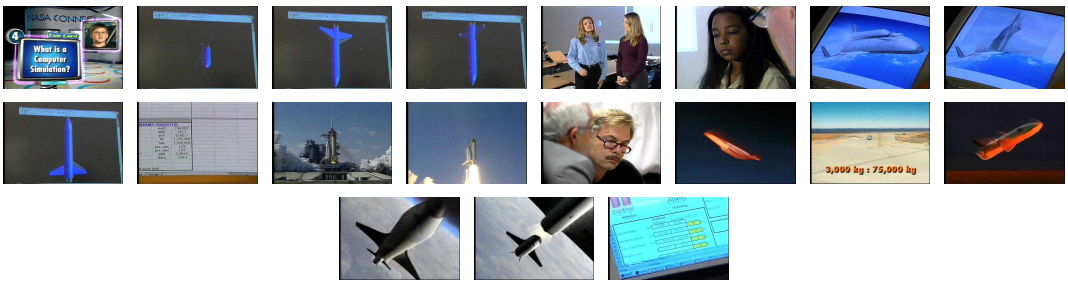
Figura A.6: Resumos gerados pelas diferentes abordagens para o vídeo *Drag Activity Part One (v06)*.

(a) Open Video. PMO = 3,3.

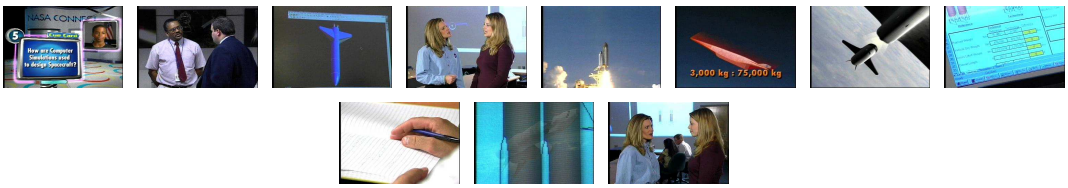


(b) Método Proposto. PMO = 3,3.

Figura A.7: Resumos gerados pelas diferentes abordagens para o vídeo *Model Testing (v07)*.




(a) Open Video. PMO = 3,8.



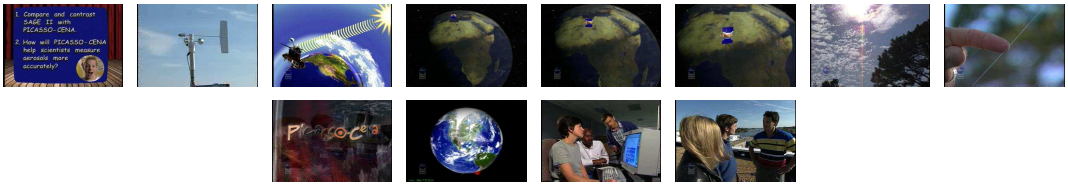

(b) Método Proposto. PMO = 3,6.

Figura A.8: Resumos gerados pelas diferentes abordagens para o vídeo *Computer Simulation (v08)*.


(a) Open Video. PMO = 3,4.


(b) Método Proposto. PMO = 3,7.

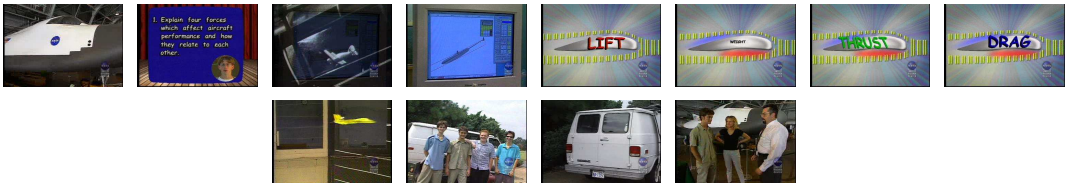
Figura A.9: Resumos gerados pelas diferentes abordagens para o vídeo *Aerosol Measurement and Remote Sensing (v09)*.



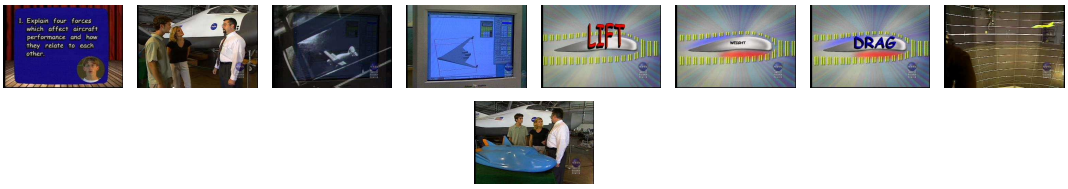
(a) Open Video. PMO = 3,8.

(b) Método Proposto. PMO = 3,6.

Figura A.10: Resumos gerados pelas diferentes abordagens para o vídeo *SAGE II and Picasso-Cena (v10)*.

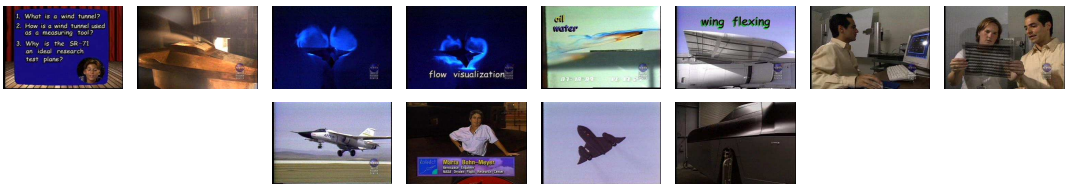


(a) Open Video. PMO = 3,8.

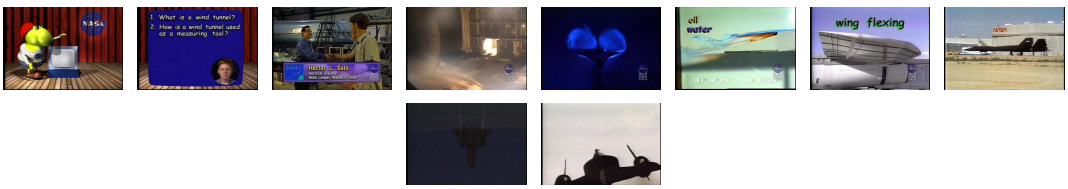


(b) Método Proposto. PMO = 4,1.

Figura A.11: Resumos gerados pelas diferentes abordagens para o vídeo *Aerodynamic Forces (v11)*.

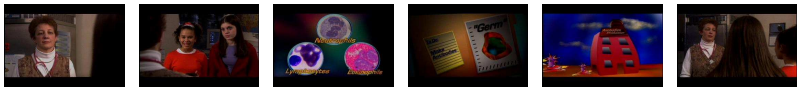


(a) Open Video. PMO = 3,7.

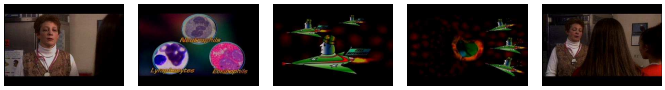


(b) Método Proposto. PMO = 3,6.

Figura A.12: Resumos gerados pelas diferentes abordagens para o vídeo *Wind Tunnels (v12)*.



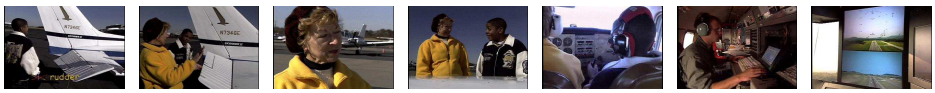
(a) Open Video. PMO = 3,7.



(b) Método Proposto. PMO = 4,0.

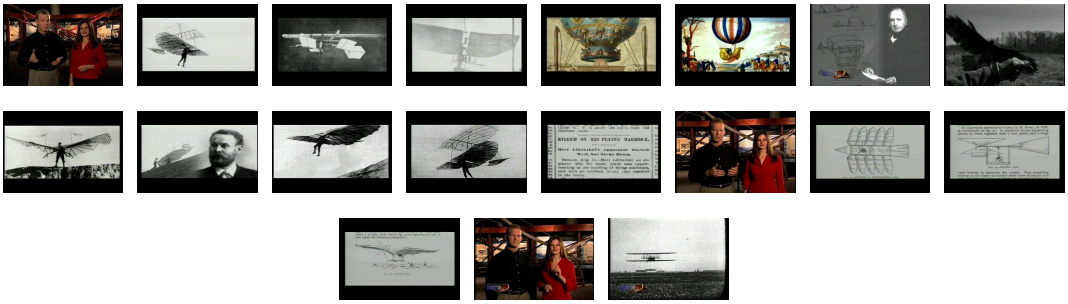
Figura A.13: Resumos gerados pelas diferentes abordagens para o vídeo *Immune System (v13)*.

(a) Open Video. PMO = 3,9.

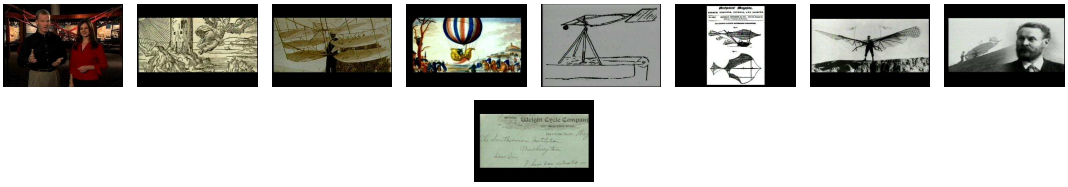


(b) Método Proposto. PMO = 3,5.

Figura A.14: Resumos gerados pelas diferentes abordagens para o vídeo *Flying a Plane (v14)*.



(a) Open Video. PMO = 3,7.



(b) Método Proposto. PMO = 4,0.

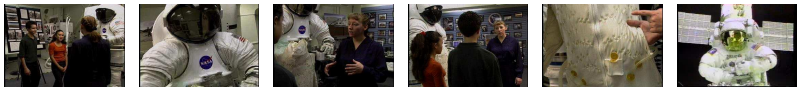
Figura A.15: Resumos gerados pelas diferentes abordagens para o vídeo *Flight Pioneers (v15)*.

(a) Open Video. PMO = 3,5.

(b) Método Proposto. PMO = 3,5.

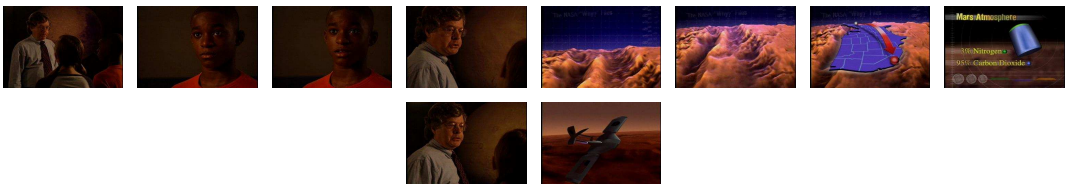
Figura A.16: Resumos gerados pelas diferentes abordagens para o vídeo *Astronauts in Space (v16)*.

(a) Open Video. PMO = 3,7.

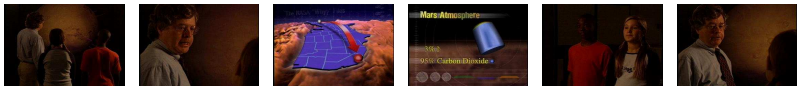


(b) Método Proposto. PMO = 4,0.

Figura A.17: Resumos gerados pelas diferentes abordagens para o vídeo *Space Suits (v17)*.



(a) Open Video. PMO = 3,8.



(b) Método Proposto. PMO = 3,5.

Figura A.18: Resumos gerados pelas diferentes abordagens para o vídeo *The Red Planet (v18)*.


(a) Open Video. PMO = 3,3.

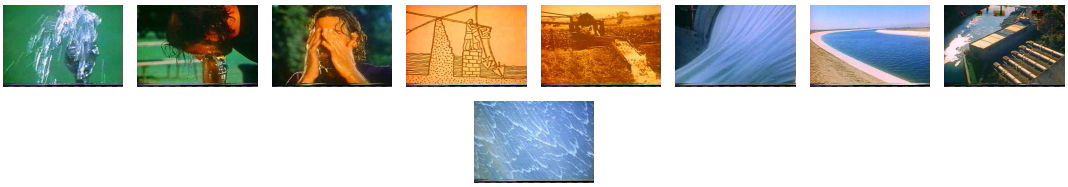

(b) Método Proposto. PMO = 3,6.

Figura A.19: Resumos gerados pelas diferentes abordagens para o vídeo *Moon Phases (v19)*.

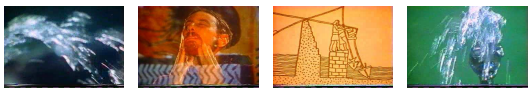
(a) Open Video. PMO = 3,0.


(b) Método Proposto. PMO = 3,6.

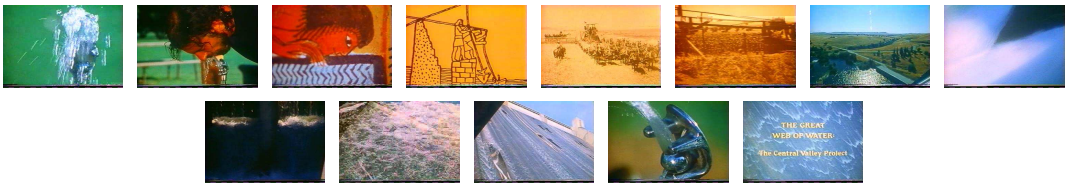
Figura A.20: Resumos gerados pelas diferentes abordagens para o vídeo *Oil Clean Up (v20)*.



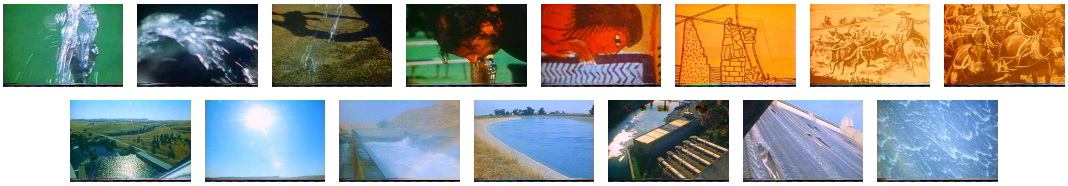
(a) Open Video. $\text{CRU}_{acerto} = 0{,}55$, $\text{CRU}_{erro} = 0{,}73$.



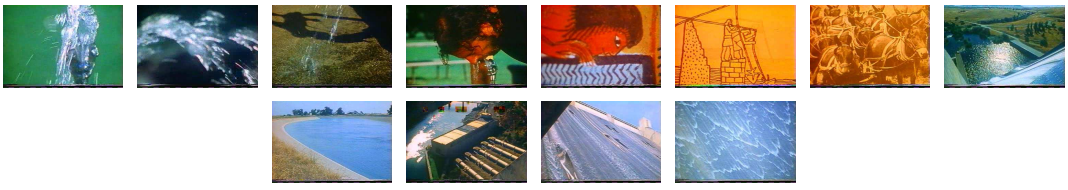
(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}20$, $\text{CRU}_{erro} = 0{,}37$.




(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}83$, $\text{CRU}_{erro} = 1{,}02$.



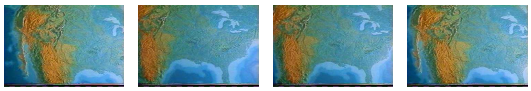
(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}79$, $\text{CRU}_{erro} = 1{,}35$.



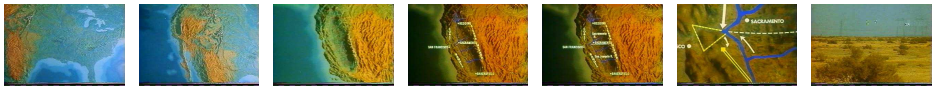
(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}75$, $\text{CRU}_{erro} = 0{,}96$.

Figura A.21: Resumos gerados pelas diferentes abordagens para o vídeo *The Great Web of Water, segment 01 (v21)*.

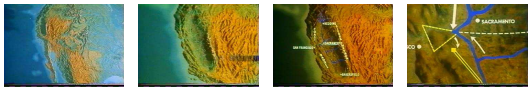
(a) Open Video. $\text{CRU}_{acerto} = 0{,}83$, $\text{CRU}_{erro} = 0{,}40$.



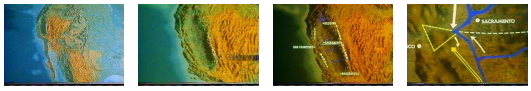
(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}17$, $\text{CRU}_{erro} = 0{,}82$.



(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}73$.

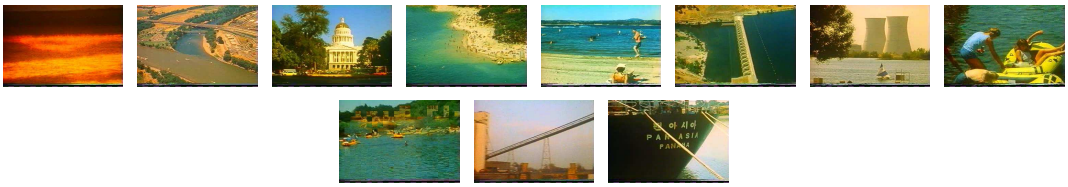


(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}56$, $\text{CRU}_{erro} = 0{,}43$.



(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}56$, $\text{CRU}_{erro} = 0{,}43$.

Figura A.22: Resumos gerados pelas diferentes abordagens para o vídeo *The Great Web of Water, segment 02 (v22).*



(a) Open Video. $\text{CRU}_{acerto} = 0{,}78$, $\text{CRU}_{erro} = 0{,}58$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}57$, $\text{CRU}_{erro} = 0{,}30$.

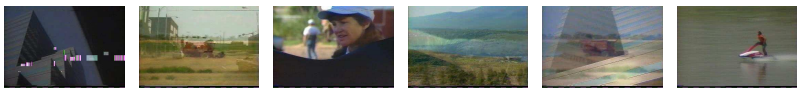
(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}52$, $\text{CRU}_{erro} = 0{,}35$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}98$, $\text{CRU}_{erro} = 0{,}88$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}93$, $\text{CRU}_{erro} = 0{,}68$.

Figura A.23: Resumos gerados pelas diferentes abordagens para o vídeo *The Great Web of Water, segment 07 (v23)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}75$, $\text{CRU}_{erro} = 0{,}39$.



(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}18$, $\text{CRU}_{erro} = 0{,}50$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}57$, $\text{CRU}_{erro} = 0{,}57$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}89$, $\text{CRU}_{erro} = 0{,}36$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}72$, $\text{CRU}_{erro} = 0{,}30$.

Figura A.24: Resumos gerados pelas diferentes abordagens para o vídeo *A New Horizon, segment 01 (v24)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}33$, $\text{CRU}_{erro} = 0{,}09$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}15$, $\text{CRU}_{erro} = 0{,}18$.

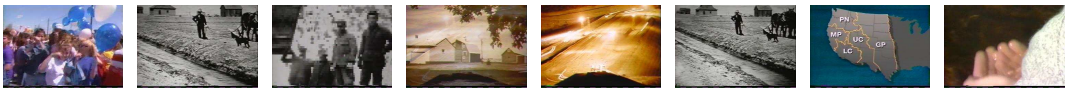

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}38$, $\text{CRU}_{erro} = 0{,}12$.


(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}66$, $\text{CRU}_{erro} = 0{,}17$.


(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}61$, $\text{CRU}_{erro} = 0{,}14$.

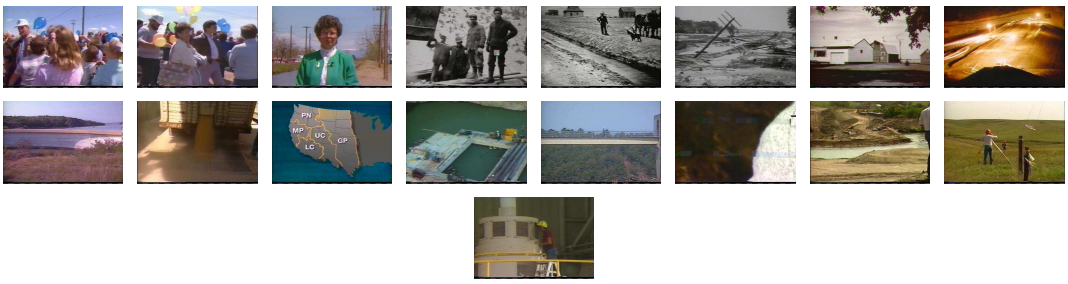
Figura A.25: Resumos gerados pelas diferentes abordagens para o vídeo *A New Horizon, segment 02 (v25)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}13$, $\text{CRU}_{erro} = 0{,}14$.



(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}38$, $\text{CRU}_{erro} = 0{,}08$.

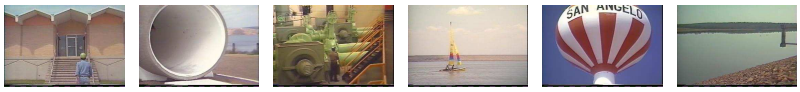
(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}66$, $\text{CRU}_{erro} = 0{,}65$.



(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}71$, $\text{CRU}_{erro} = 0{,}40$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}53$, $\text{CRU}_{erro} = 0{,}25$.

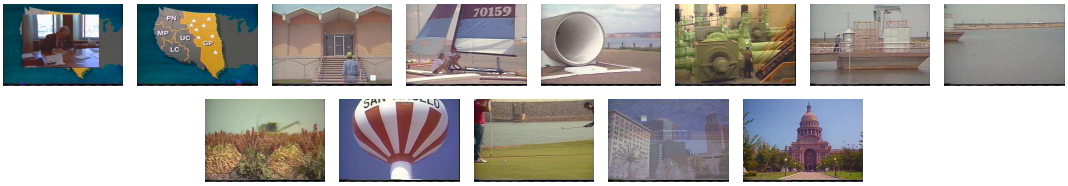
Figura A.26: Resumos gerados pelas diferentes abordagens para o vídeo *A New Horizon, segment 03 (v26)*.



(a) Open Video. $\text{CRU}_{acerto} = 0{,}25$, $\text{CRU}_{erro} = 0{,}55$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}48$, $\text{CRU}_{erro} = 0{,}45$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}44$, $\text{CRU}_{erro} = 0{,}89$.



(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}78$, $\text{CRU}_{erro} = 0{,}94$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}60$, $\text{CRU}_{erro} = 0{,}59$.

Figura A.27: Resumos gerados pelas diferentes abordagens para o vídeo *A New Horizon, segment 04 (v27)*.


(a) Open Video. $\text{CRU}_{acerto} = 0{,}63$, $\text{CRU}_{erro} = 0{,}12$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}24$, $\text{CRU}_{erro} = 0{,}08$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}43$, $\text{CRU}_{erro} = 0{,}15$.

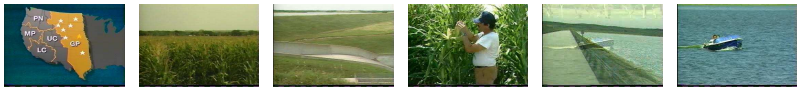
(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}82$, $\text{CRU}_{erro} = 0{,}19$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}68$, $\text{CRU}_{erro} = 0{,}17$.

Figura A.28: Resumos gerados pelas diferentes abordagens para o vídeo *A New Horizon, segment 05 (v28)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}48$, $\text{CRU}_{erro} = 0{,}31$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}56$, $\text{CRU}_{erro} = 0{,}55$.



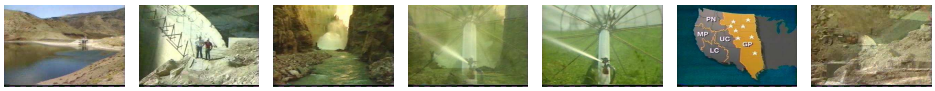
(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}54$, $\text{CRU}_{erro} = 0{,}41$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}75$, $\text{CRU}_{erro} = 0{,}68$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}75$, $\text{CRU}_{erro} = 0{,}52$.

Figura A.29: Resumos gerados pelas diferentes abordagens para o vídeo *A New Horizon, segment 06 (v29)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}62$, $\text{CRU}_{erro} = 0{,}11$.



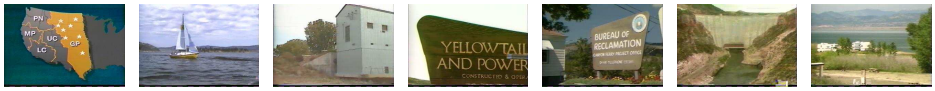
(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}58$, $\text{CRU}_{erro} = 0{,}16$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}60$, $\text{CRU}_{erro} = 0{,}13$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}75$, $\text{CRU}_{erro} = 0{,}09$.

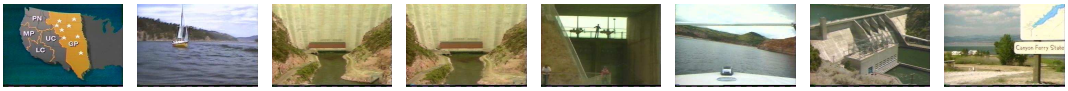
(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}75$, $\text{CRU}_{erro} = 0{,}09$.

Figura A.30: Resumos gerados pelas diferentes abordagens para o vídeo *A New Horizon, segment 08 (v30)*.



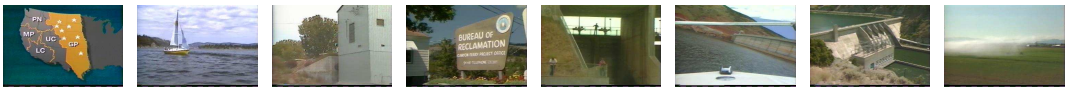
(a) Open Video. $\text{CRU}_{acerto} = 0{,}64$, $\text{CRU}_{erro} = 0{,}16$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}53$, $\text{CRU}_{erro} = 0{,}16$.



(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}60$, $\text{CRU}_{erro} = 0{,}32$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}84$, $\text{CRU}_{erro} = 0{,}42$.



(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}63$, $\text{CRU}_{erro} = 0{,}28$.

Figura A.31: Resumos gerados pelas diferentes abordagens para o vídeo *A New Horizon, segment 10 (v31)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}53$, $\text{CRU}_{erro} = 0{,}45$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}18$, $\text{CRU}_{erro} = 0{,}19$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}62$, $\text{CRU}_{erro} = 0{,}60$.

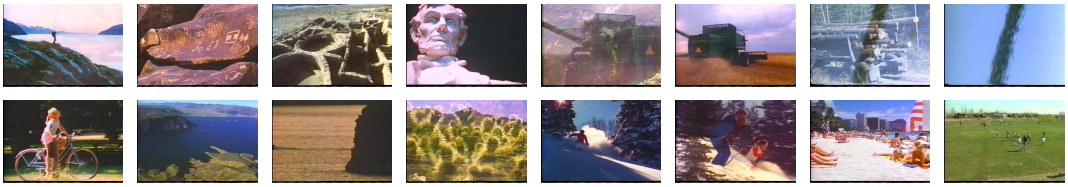
(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}81$, $\text{CRU}_{erro} = 0{,}54$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}77$, $\text{CRU}_{erro} = 0{,}58$.

Figura A.32: Resumos gerados pelas diferentes abordagens para o vídeo *Take Pride in America, segment 01 (v32)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}56$, $\text{CRU}_{erro} = 0{,}36$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}22$, $\text{CRU}_{erro} = 0{,}20$.



(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}55$, $\text{CRU}_{erro} = 0{,}59$.

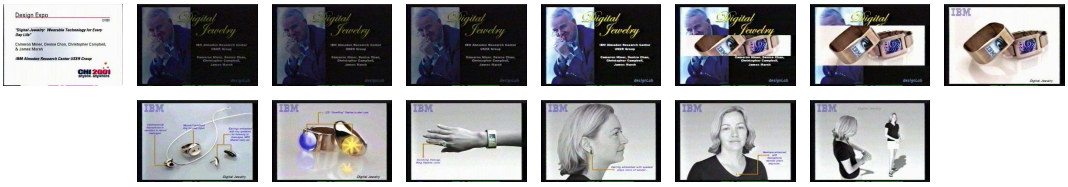
(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}76$, $\text{CRU}_{erro} = 0{,}67$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}51$, $\text{CRU}_{erro} = 0{,}41$.

Figura A.33: Resumos gerados pelas diferentes abordagens para o vídeo *Take Pride in America, segment 03 (v33)*.

(a) Open Video. $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}70$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}60$, $\text{CRU}_{erro} = 0{,}10$.

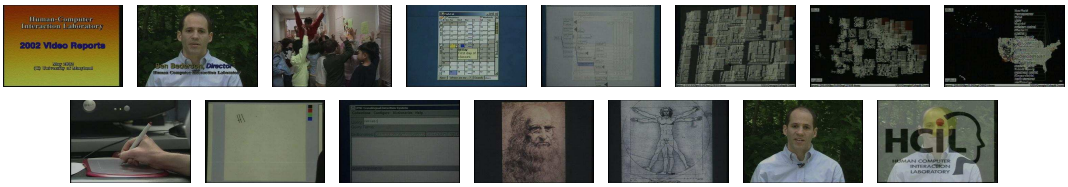


(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}94$, $\text{CRU}_{erro} = 0{,}46$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}90$, $\text{CRU}_{erro} = 0{,}00$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}70$, $\text{CRU}_{erro} = 0{,}00$.

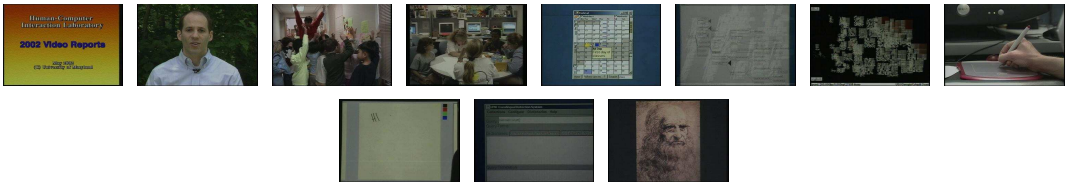
Figura A.34: Resumos gerados pelas diferentes abordagens para o vídeo *Digital Jewelry: Wearable Technology for Every Day Life (v34)*.



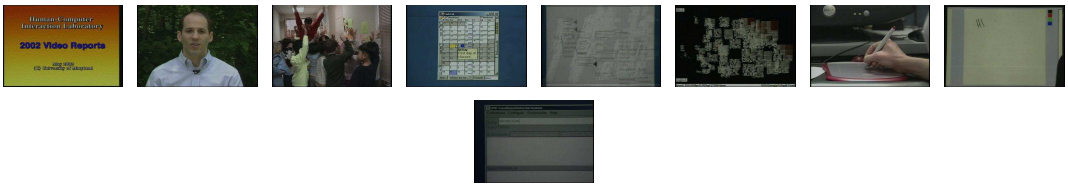
(a) Open Video. $\text{CRU}_{acerto} = 0{,}80$, $\text{CRU}_{erro} = 0{,}73$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}30$, $\text{CRU}_{erro} = 0{,}31$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}42$, $\text{CRU}_{erro} = 0{,}29$.



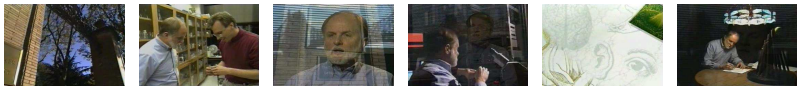
(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}87$, $\text{CRU}_{erro} = 0{,}26$.



(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}69$, $\text{CRU}_{erro} = 0{,}22$.

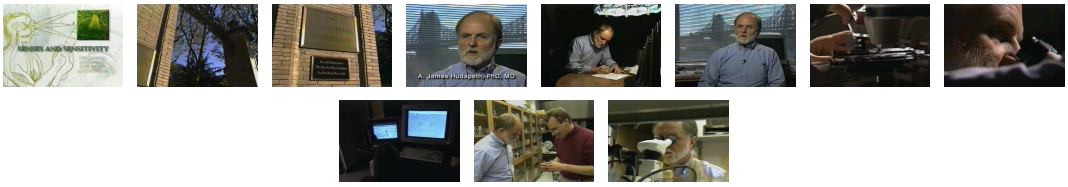
Figura A.35: Resumos gerados pelas diferentes abordagens para o vídeo *HCIL Symposium 2002 – Introduction, segment 01 (v35).*

(a) Open Video. $\text{CRU}_{acerto} = 0{,}89$, $\text{CRU}_{erro} = 1{,}17$.

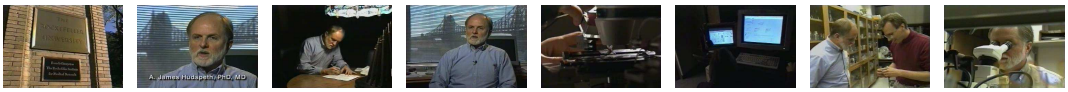


(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}29$, $\text{CRU}_{erro} = 0{,}83$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}85$, $\text{CRU}_{erro} = 1{,}76$.



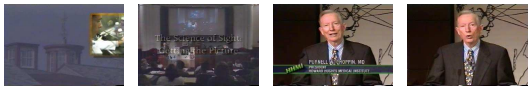
(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}77$, $\text{CRU}_{erro} = 1{,}28$.



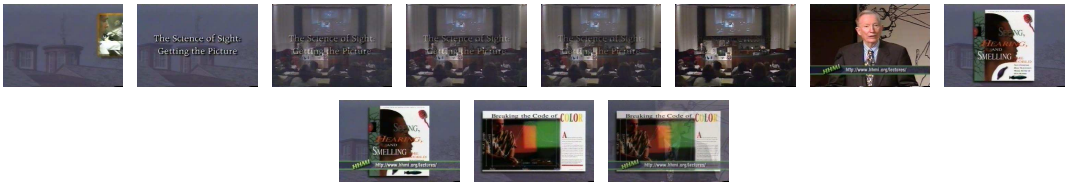
(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}63$, $\text{CRU}_{erro} = 0{,}87$.

Figura A.36: Resumos gerados pelas diferentes abordagens para o vídeo *Senses And Sensitivity, Introduction to Lecture 1 presenter (v36).*

(a) Open Video. $\text{CRU}_{acerto} = 0{,}60$, $\text{CRU}_{erro} = 0{,}20$.



(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}60$, $\text{CRU}_{erro} = 0{,}20$.



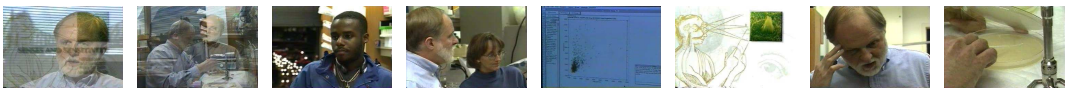
(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 1{,}20$.

(d) VSUMM1. $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}00$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}60$, $\text{CRU}_{erro} = 0{,}00$.

Figura A.37: Resumos gerados pelas diferentes abordagens para o vídeo *Senses And Sensitivity, Introduction to Lecture 2 (v37)*.

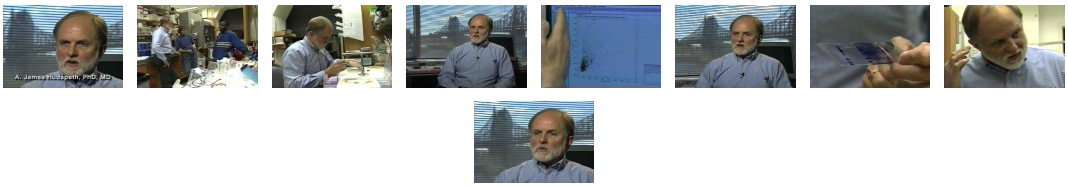
(a) Open Video. $\text{CRU}_{acerto} = 0{,}60$, $\text{CRU}_{erro} = 0{,}30$.



(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}38$, $\text{CRU}_{erro} = 0{,}34$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}79$, $\text{CRU}_{erro} = 0{,}57$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}84$, $\text{CRU}_{erro} = 0{,}43$.



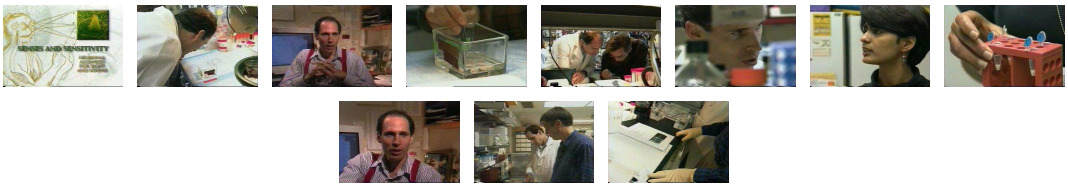
(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}55$, $\text{CRU}_{erro} = 0{,}26$.

Figura A.38: Resumos gerados pelas diferentes abordagens para o vídeo *Senses And Sensitivity, Introduction to Lecture 3 presenter (v38)*.

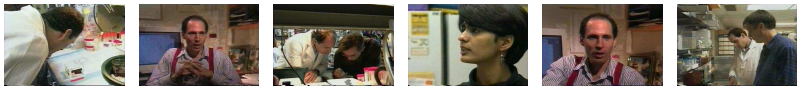
(a) Open Video. $\text{CRU}_{acerto} = 0{,}77$, $\text{CRU}_{erro} = 0{,}30$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}44$, $\text{CRU}_{erro} = 0{,}05$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}93$, $\text{CRU}_{erro} = 0{,}46$.

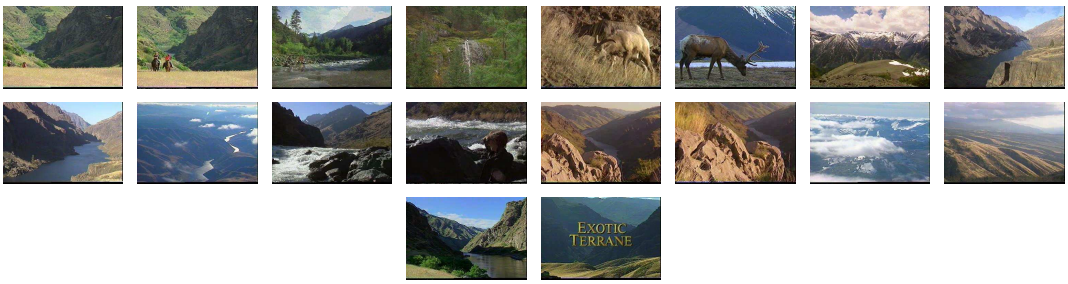


(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}83$, $\text{CRU}_{erro} = 0{,}09$.



(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}46$, $\text{CRU}_{erro} = 0{,}13$.

Figura A.39: Resumos gerados pelas diferentes abordagens para o vídeo *Senses And Sensitivity, Introduction to Lecture 4 presenter (v39).*




(a) Open Video. $\text{CRU}_{acerto} = 0{,}98$, $\text{CRU}_{erro} = 1{,}39$.

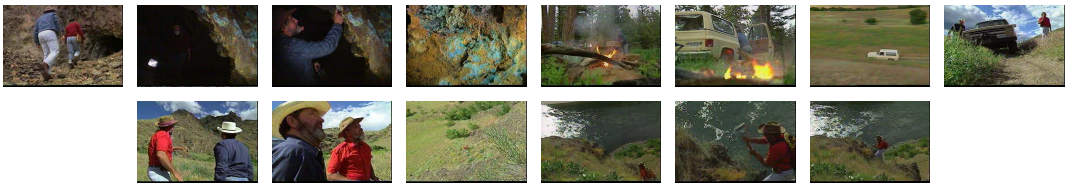
(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}48$, $\text{CRU}_{erro} = 0{,}32$.


(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}66$, $\text{CRU}_{erro} = 0{,}53$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}82$, $\text{CRU}_{erro} = 0{,}89$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}69$, $\text{CRU}_{erro} = 0{,}63$

Figura A.40: Resumos gerados pelas diferentes abordagens para o vídeo *Exotic Terrane, segment 01 (v40)*.



(a) Open Video. $\text{CRU}_{acerto} = 0{,}98$, $\text{CRU}_{erro} = 0{,}81$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}47$, $\text{CRU}_{erro} = 0{,}68$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}62$, $\text{CRU}_{erro} = 0{,}53$.

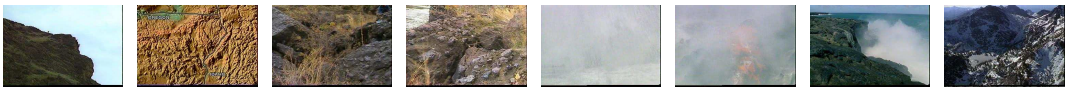
(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}89$, $\text{CRU}_{erro} = 0{,}39$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}71$, $\text{CRU}_{erro} = 0{,}19$.

Figura A.41: Resumos gerados pelas diferentes abordagens para o vídeo *Exotic Terrane, segment 02 (v41)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}93$, $\text{CRU}_{erro} = 0{,}81$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}54$, $\text{CRU}_{erro} = 0{,}08$.

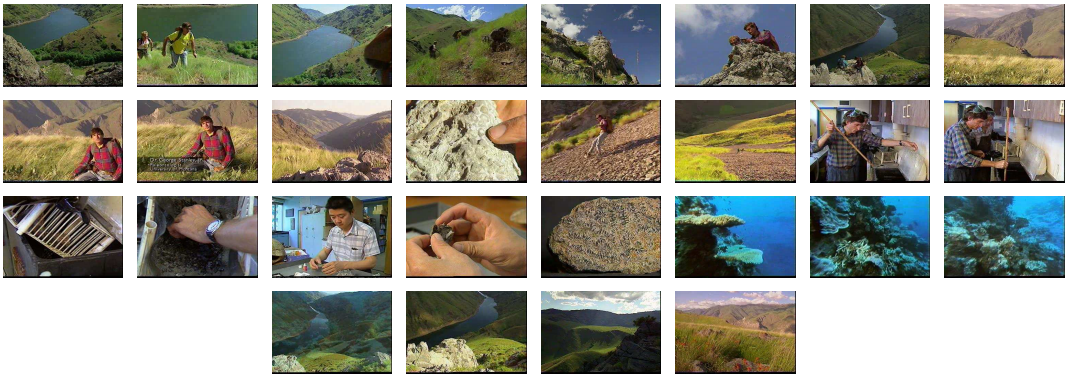


(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}66$, $\text{CRU}_{erro} = 0{,}33$.

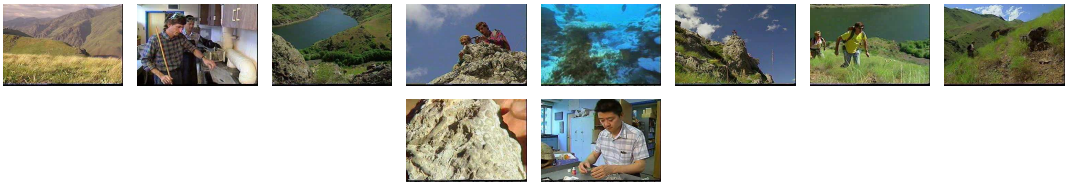
(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}74$, $\text{CRU}_{erro} = 0{,}13$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}42$, $\text{CRU}_{erro} = 0{,}08$.

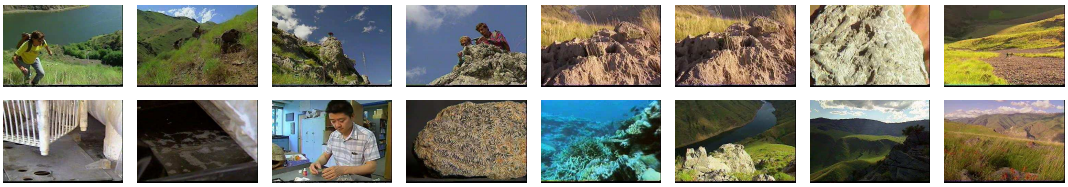
Figura A.42: Resumos gerados pelas diferentes abordagens para o vídeo *Exotic Terrane, segment 03 (v42)*.



(a) Open Video. $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 1{,}11$.



(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}59$, $\text{CRU}_{erro} = 0{,}17$.



(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}88$, $\text{CRU}_{erro} = 0{,}33$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}93$, $\text{CRU}_{erro} = 0{,}28$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}63$, $\text{CRU}_{erro} = 0{,}13$.

Figura A.43: Resumos gerados pelas diferentes abordagens para o vídeo *Exotic Terrane, segment 04 (v43)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}91$, $\text{CRU}_{erro} = 0{,}22$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}45$, $\text{CRU}_{erro} = 0{,}11$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}76$, $\text{CRU}_{erro} = 0{,}25$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}93$, $\text{CRU}_{erro} = 0{,}30$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}76$, $\text{CRU}_{erro} = 0{,}14$.

Figura A.44: Resumos gerados pelas diferentes abordagens para o vídeo *Exotic Terrane, segment 06 (v44)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}82$, $\text{CRU}_{erro} = 1{,}14$.

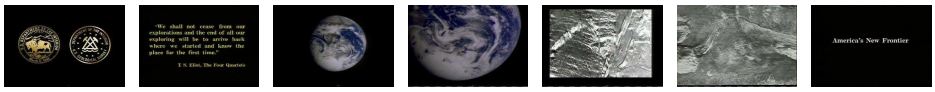
(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}25$, $\text{CRU}_{erro} = 0{,}24$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}78$, $\text{CRU}_{erro} = 0{,}52$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}84$, $\text{CRU}_{erro} = 0{,}30$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}73$, $\text{CRU}_{erro} = 0{,}25$.

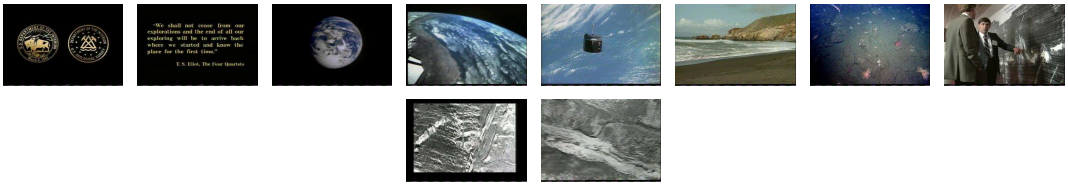
Figura A.45: Resumos gerados pelas diferentes abordagens para o vídeo *Exotic Terrane, segment 08 (v45)*.



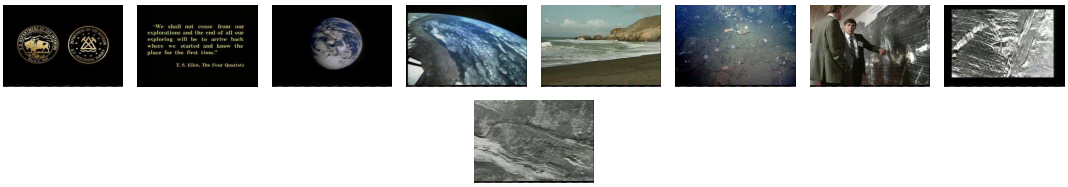

(a) Open Video. $\text{CRU}_{acerto} = 0{,}54$, $\text{CRU}_{erro} = 0{,}42$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}66$, $\text{CRU}_{erro} = 0{,}44$.


(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}51$, $\text{CRU}_{erro} = 1{,}00$.



(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}92$, $\text{CRU}_{erro} = 0{,}46$.



(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}86$, $\text{CRU}_{erro} = 0{,}37$.

Figura A.46: Resumos gerados pelas diferentes abordagens para o vídeo *America's New Frontier, segment 01 (v46).*

(a) Open Video. $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}40$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}96$, $\text{CRU}_{erro} = 0{,}04$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}40$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}96$, $\text{CRU}_{erro} = 0{,}04$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}76$, $\text{CRU}_{erro} = 0{,}04$.

Figura A.47: Resumos gerados pelas diferentes abordagens para o vídeo *America's New Frontier, segment 03 (v47)*.

(a) Open Video. $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}93$.

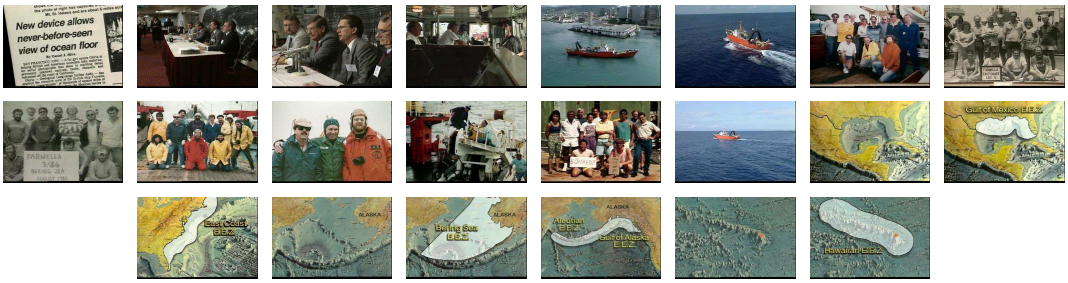
(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}92$, $\text{CRU}_{erro} = 0{,}12$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}79$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}92$, $\text{CRU}_{erro} = 0{,}12$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}92$, $\text{CRU}_{erro} = 0{,}12$.

Figura A.48: Resumos gerados pelas diferentes abordagens para o vídeo *America's New Frontier, segment 04 (v48)*.




(a) Open Video. $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}71$.


(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}61$, $\text{CRU}_{erro} = 0{,}17$.

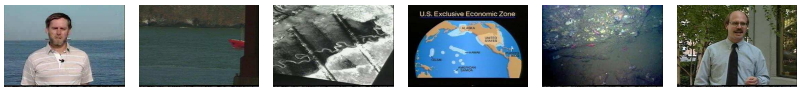

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}58$, $\text{CRU}_{erro} = 0{,}35$.


(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}93$, $\text{CRU}_{erro} = 0{,}08$.


(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}67$, $\text{CRU}_{erro} = 0{,}03$.

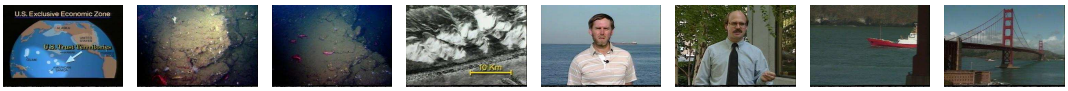
Figura A.49: Resumos gerados pelas diferentes abordagens para o vídeo *America's New Frontier, segment 07 (v49)*.

(a) Open Video. $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 1{,}18$.



(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}65$, $\text{CRU}_{erro} = 0{,}22$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 1{,}33$.



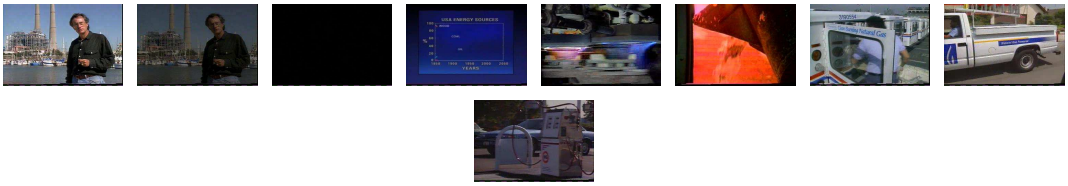
(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}89$, $\text{CRU}_{erro} = 0{,}27$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}85$, $\text{CRU}_{erro} = 0{,}17$.

Figura A.50: Resumos gerados pelas diferentes abordagens para o vídeo *America's New Frontier, segment 10 (v50)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}85$, $\text{CRU}_{erro} = 1{,}25$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}72$, $\text{CRU}_{erro} = 0{,}40$.

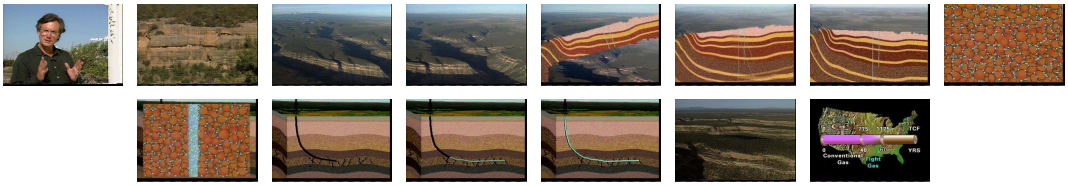


(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}74$, $\text{CRU}_{erro} = 0{,}52$.

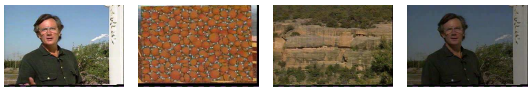
(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}88$, $\text{CRU}_{erro} = 0{,}24$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}50$, $\text{CRU}_{erro} = 0{,}07$.

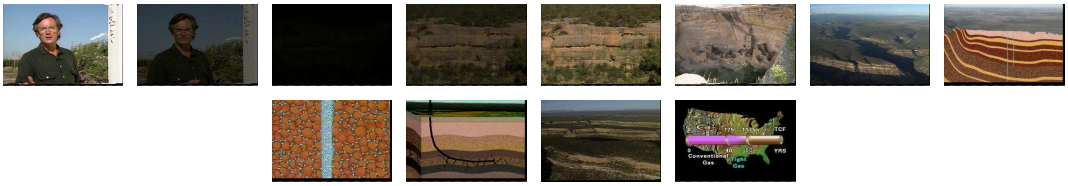
Figura A.51: Resumos gerados pelas diferentes abordagens para o vídeo *The Future of Energy Gases, segment 03 (v51).*



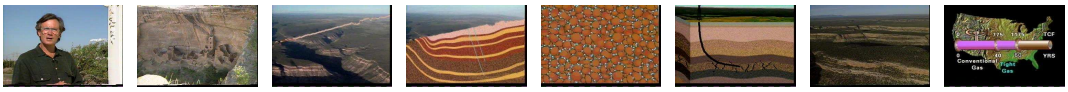
(a) Open Video. $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}76$.



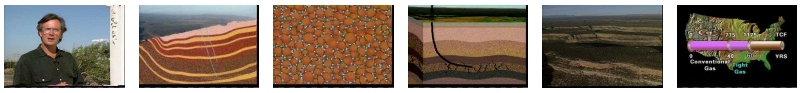
(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}41$, $\text{CRU}_{erro} = 0{,}10$.



(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}51$.



(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}95$, $\text{CRU}_{erro} = 0{,}05$.

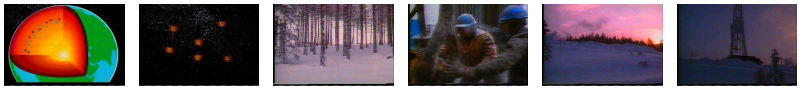


(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}75$, $\text{CRU}_{erro} = 0{,}00$.

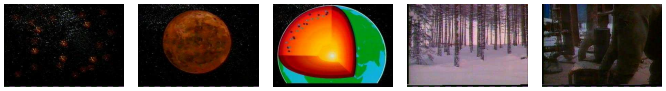
Figura A.52: Resumos gerados pelas diferentes abordagens para o vídeo *The Future of Energy Gases, segment 05 (v52)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}03$, $\text{CRU}_{erro} = 3{,}47$.

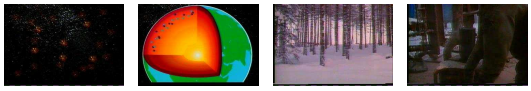
(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}55$, $\text{CRU}_{erro} = 0{,}70$.



(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}67$ , $\text{CRU}_{erro} = 0{,}83$.



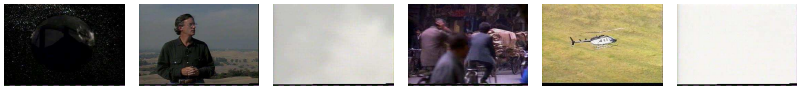
(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}50$, $\text{CRU}_{erro} = 0{,}75$.



(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}42$, $\text{CRU}_{erro} = 0{,}58$.

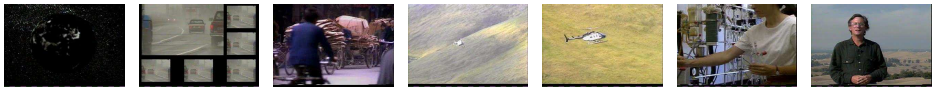
Figura A.53: Resumos gerados pelas diferentes abordagens para o vídeo *The Future of Energy Gases, segment 09 (v53).*

(a) Open Video. $\text{CRU}_{acerto} = 0{,}74$, $\text{CRU}_{erro} = 0{,}46$.

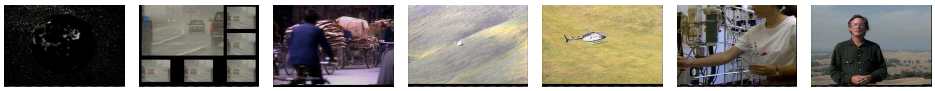


(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}51$, $\text{CRU}_{erro} = 0{,}40$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}89$, $\text{CRU}_{erro} = 0{,}47$.



(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}85$, $\text{CRU}_{erro} = 0{,}21$.

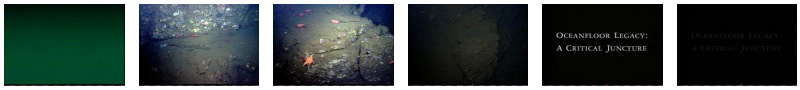


(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}85$, $\text{CRU}_{erro} = 0{,}21$.

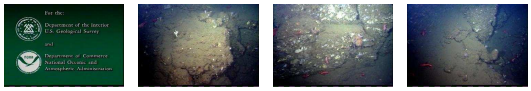
Figura A.54: Resumos gerados pelas diferentes abordagens para o vídeo *The Future of Energy Gases, segment 12 (v54)*.


(a) Open Video. $\text{CRU}_{acerto} = 0{,}23$, $\text{CRU}_{erro} = 0{,}36$.

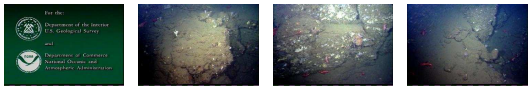
(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}36$, $\text{CRU}_{erro} = 0{,}53$.




(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}61$, $\text{CRU}_{erro} = 1{,}17$.



(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}75$, $\text{CRU}_{erro} = 0{,}43$.



(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}75$, $\text{CRU}_{erro} = 0{,}43$.

Figura A.55: Resumos gerados pelas diferentes abordagens para o vídeo *Oceanfloor Legacy, segment 01 (v55).*

(a) Open Video. $\text{CRU}_{acerto} = 0{,}69$, $\text{CRU}_{erro} = 0{,}38$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}45$, $\text{CRU}_{erro} = 0{,}44$.

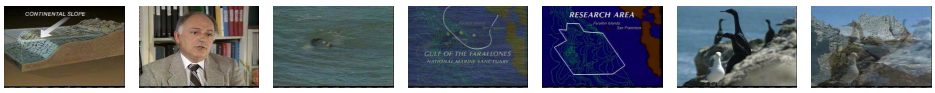
(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}61$, $\text{CRU}_{erro} = 0{,}64$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}96$, $\text{CRU}_{erro} = 0{,}48$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}69$, $\text{CRU}_{erro} = 0{,}20$.

Figura A.56: Resumos gerados pelas diferentes abordagens para o vídeo *Oceanfloor Legacy, segment 02 (v56)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}73$, $\text{CRU}_{erro} = 0{,}31$.



(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}59$, $\text{CRU}_{erro} = 0{,}45$.

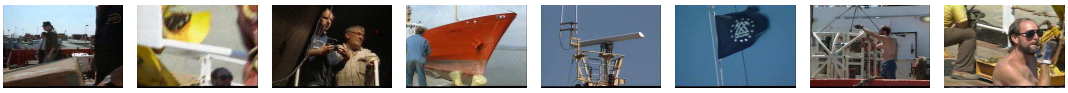
(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}88$, $\text{CRU}_{erro} = 0{,}76$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}90$, $\text{CRU}_{erro} = 0{,}44$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}86$, $\text{CRU}_{erro} = 0{,}34$.

Figura A.57: Resumos gerados pelas diferentes abordagens para o vídeo *Oceanfloor Legacy, segment 04 (v57).*

(a) Open Video. $\text{CRU}_{acerto} = 0{,}44$, $\text{CRU}_{erro} = 0{,}09$.



(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}65$, $\text{CRU}_{erro} = 0{,}21$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}74$, $\text{CRU}_{erro} = 0{,}33$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}90$, $\text{CRU}_{erro} = 0{,}49$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}75$, $\text{CRU}_{erro} = 0{,}22$.

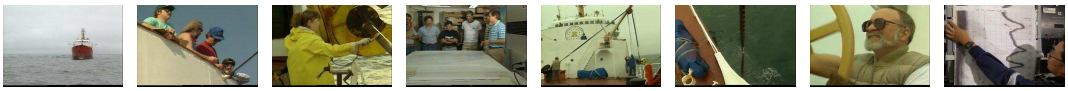
Figura A.58: Resumos gerados pelas diferentes abordagens para o vídeo *Oceanfloor Legacy, segment 08 (v58)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}72$, $\text{CRU}_{erro} = 0{,}03$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}69$, $\text{CRU}_{erro} = 0{,}16$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}70$, $\text{CRU}_{erro} = 0{,}16$.

(d) VSUMM1. $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}28$.



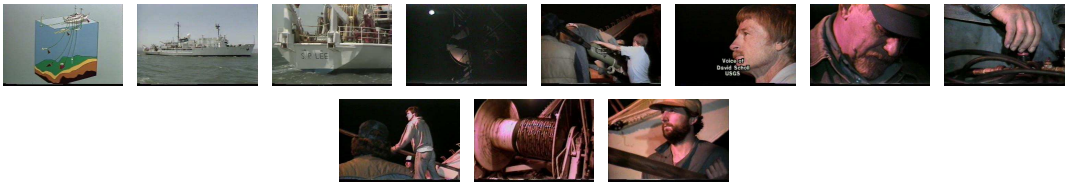
(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}81$, $\text{CRU}_{erro} = 0{,}05$.

Figura A.59: Resumos gerados pelas diferentes abordagens para o vídeo *Oceanfloor Legacy, segment 09 (v59)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}83$, $\text{CRU}_{erro} = 0{,}41$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}64$, $\text{CRU}_{erro} = 0{,}36$.

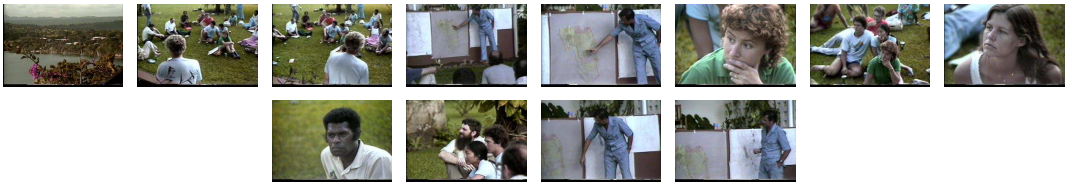
(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}55$, $\text{CRU}_{erro} = 0{,}32$.



(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}93$, $\text{CRU}_{erro} = 0{,}44$.

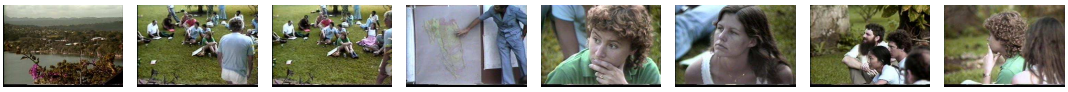
(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}81$, $\text{CRU}_{erro} = 0{,}31$.

Figura A.60: Resumos gerados pelas diferentes abordagens para o vídeo *The Voyage of the Lee, segment 05 (v60)*.



(a) Open Video. $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 1{,}62$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}81$, $\text{CRU}_{erro} = 0{,}29$.



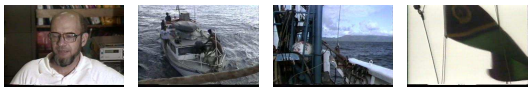
(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}89$, $\text{CRU}_{erro} = 0{,}86$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}93$, $\text{CRU}_{erro} = 0{,}60$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}74$, $\text{CRU}_{erro} = 0{,}13$.

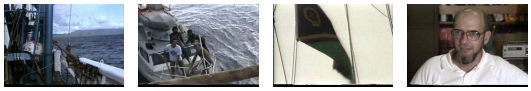
Figura A.61: Resumos gerados pelas diferentes abordagens para o vídeo *The Voyage of the Lee, segment 15 (v61)*.

(a) Open Video. $\text{CRU}_{acerto} = 0{,}78$, $\text{CRU}_{erro} = 1{,}20$.



(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}13$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}98$.

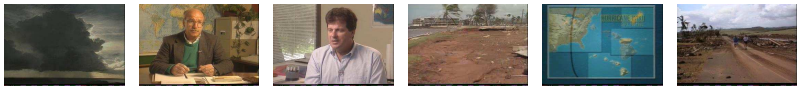


(d) VSUMM1. $\text{CRU}_{acerto} = 1{,}00$, $\text{CRU}_{erro} = 0{,}13$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}78$, $\text{CRU}_{erro} = 0{,}07$.

Figura A.62: Resumos gerados pelas diferentes abordagens para o vídeo *The Voyage of the Lee, segment 16 (v62).*

(a) Open Video. $\text{CRU}_{acerto} = 0{,}59$, $\text{CRU}_{erro} = 0{,}25$.



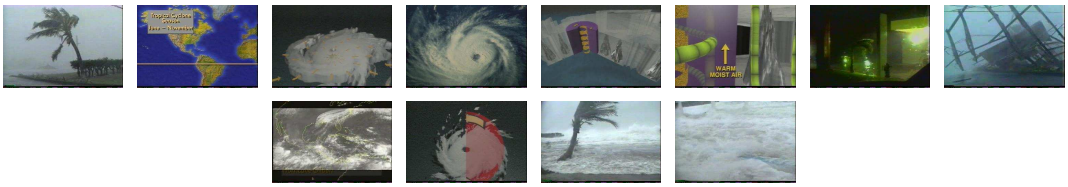
(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}64$, $\text{CRU}_{erro} = 0{,}20$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}36$, $\text{CRU}_{erro} = 0{,}61$.

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}89$, $\text{CRU}_{erro} = 0{,}36$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}70$, $\text{CRU}_{erro} = 0{,}14$.

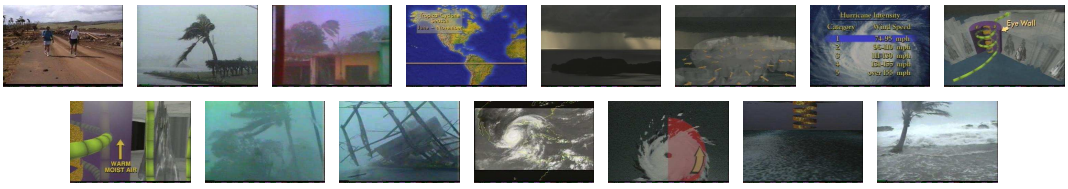
Figura A.63: Resumos gerados pelas diferentes abordagens para o vídeo *Hurricane Force – A Coastal Perspective, segment 03 (v63).*



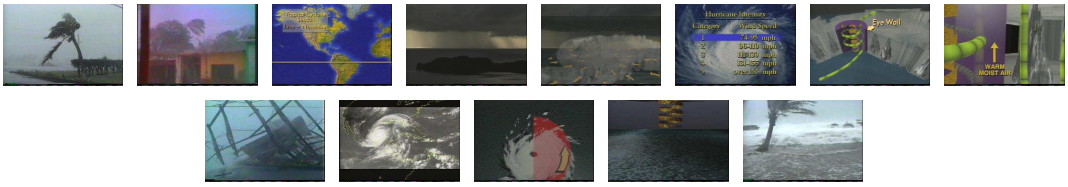
(a) Open Video. $\text{CRU}_{acerto} = 0{,}65$, $\text{CRU}_{erro} = 0{,}22$.

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}51$, $\text{CRU}_{erro} = 0{,}29$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}72$, $\text{CRU}_{erro} = 0{,}51$.

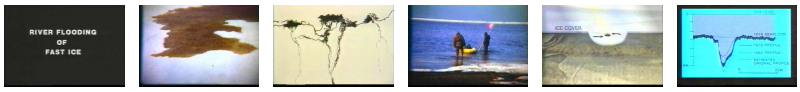


(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}76$, $\text{CRU}_{erro} = 0{,}33$.

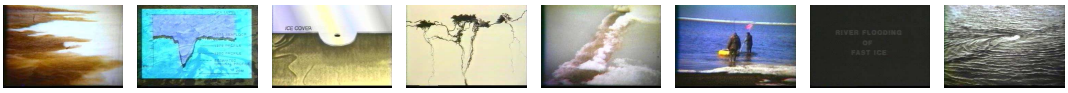


(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}67$, $\text{CRU}_{erro} = 0{,}28$.

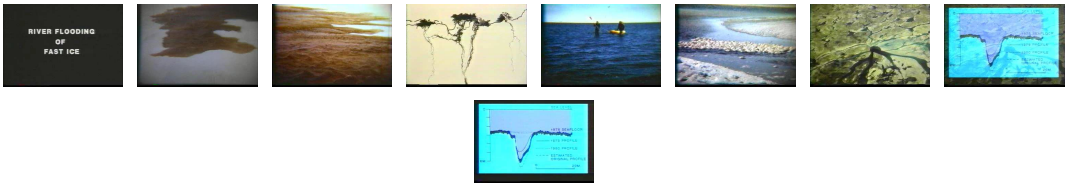
Figura A.64: Resumos gerados pelas diferentes abordagens para o vídeo *Hurricane Force – A Coastal Perspective, segment 04 (v64).*




(a) Open Video. $\text{CRU}_{acerto} = 0{,}58$, $\text{CRU}_{erro} = 0{,}22$.

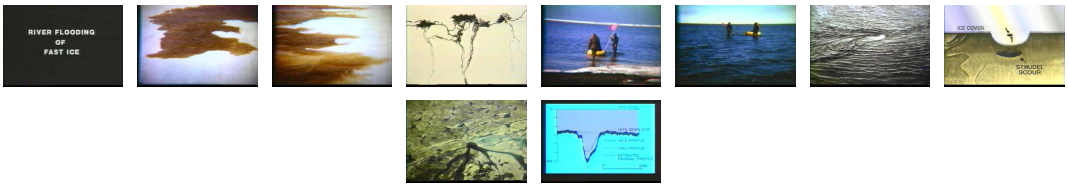


(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}61$, $\text{CRU}_{erro} = 0{,}46$.



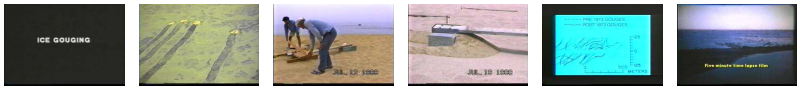

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}57$, $\text{CRU}_{erro} = 0{,}63$.


(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}98$, $\text{CRU}_{erro} = 0{,}49$.

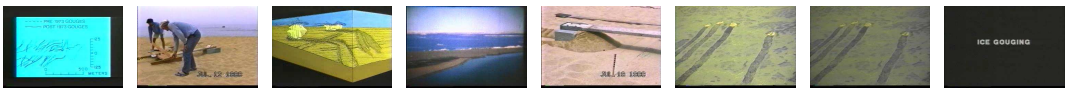



(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}91$, $\text{CRU}_{erro} = 0{,}43$.

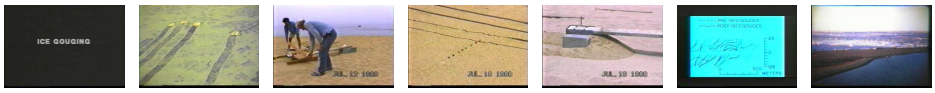
Figura A.65: Resumos gerados pelas diferentes abordagens para o vídeo *Drift Ice as a Geologic Agent, segment 03 (v65)*.




(a) Open Video. $\text{CRU}_{acerto} = 0{,}61$, $\text{CRU}_{erro} = 0{,}19$.




(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}77$, $\text{CRU}_{erro} = 0{,}30$.

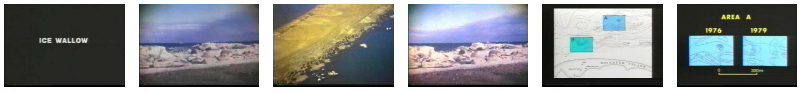



(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}68$, $\text{CRU}_{erro} = 0{,}25$.


(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}87$, $\text{CRU}_{erro} = 0{,}33$.


(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}87$, $\text{CRU}_{erro} = 0{,}33$.

Figura A.66: Resumos gerados pelas diferentes abordagens para o vídeo *Drift Ice as a Geologic Agent, segment 05 (v66)*.




(a) Open Video. $\text{CRU}_{acerto} = 0{,}69$, $\text{CRU}_{erro} = 0{,}11$.

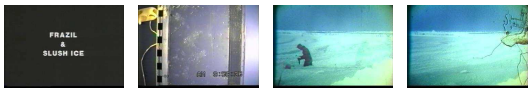

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}93$, $\text{CRU}_{erro} = 0{,}00$.

(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}79$, $\text{CRU}_{erro} = 0{,}41$.


(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}93$, $\text{CRU}_{erro} = 0{,}00$.

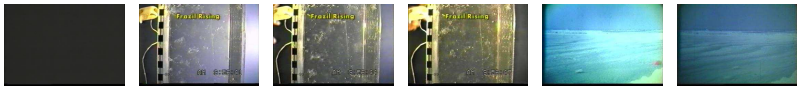

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}80$, $\text{CRU}_{erro} = 0{,}00$.

Figura A.67: Resumos gerados pelas diferentes abordagens para o vídeo *Drift Ice as a Geologic Agent, segment 06 (v67)*.

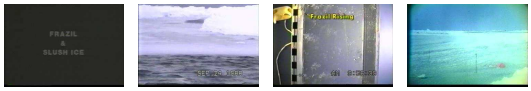



(a) Open Video. $\text{CRU}_{acerto} = 0{,}65$, $\text{CRU}_{erro} = 0{,}27$.

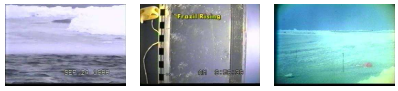

(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}69$, $\text{CRU}_{erro} = 0{,}23$.



(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}77$, $\text{CRU}_{erro} = 0{,}61$.



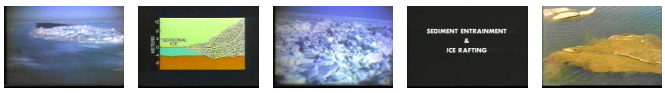
(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}88$, $\text{CRU}_{erro} = 0{,}04$.



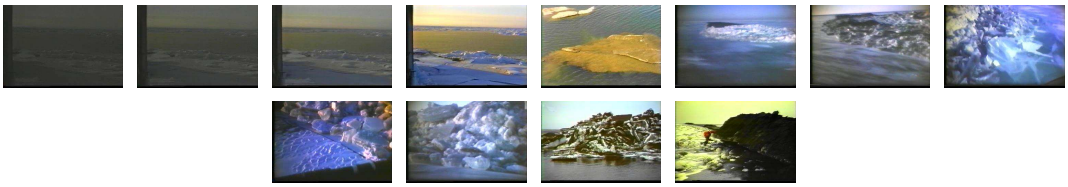
(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}65$, $\text{CRU}_{erro} = 0{,}04$.

Figura A.68: Resumos gerados pelas diferentes abordagens para o vídeo *Drift Ice as a Geologic Agent, segment 07 (v68).*

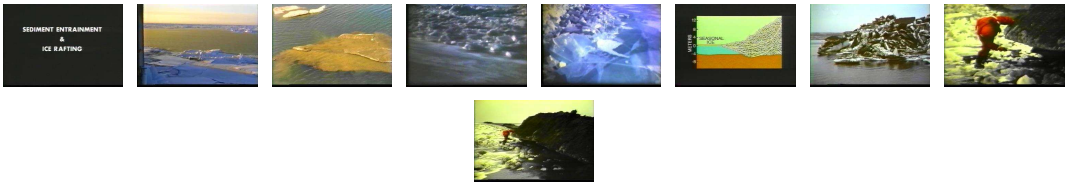

(a) Open Video. $\text{CRU}_{acerto} = 0{,}54$, $\text{CRU}_{erro} = 0{,}07$.




(b) DT (Mundur et al., 2006). $\text{CRU}_{acerto} = 0{,}49$, $\text{CRU}_{erro} = 0{,}13$.



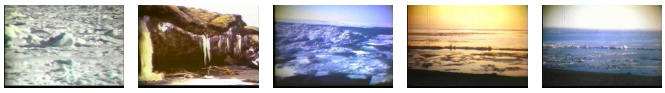
(c) VISTO (Furini et al., 2007). $\text{CRU}_{acerto} = 0{,}67$, $\text{CRU}_{erro} = 0{,}78$.



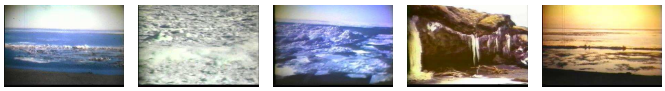

(d) VSUMM1. $\text{CRU}_{acerto} = 0{,}94$, $\text{CRU}_{erro} = 0{,}15$.

(e) VSUMM2. $\text{CRU}_{acerto} = 0{,}73$, $\text{CRU}_{erro} = 0{,}12$.

Figura A.69: Resumos gerados pelas diferentes abordagens para o vídeo *Drift Ice as a Geologic Agent, segment 08 (v69)*.



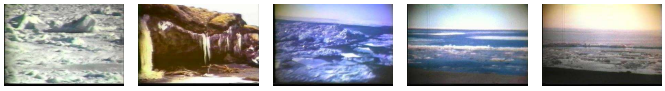
(a) Open Video. $CRU_{acerto}$ =0,96 , $CRU_{erro}$ = 0,09.



(b) DT (Mundur et al., 2006). $CRU_{acerto}$ = 0,96, $CRU_{erro}$ = 0,09.

(c) VISTO (Furini et al., 2007). $CRU_{acerto}$ = 0,84, $CRU_{erro}$ = 0,21.

(d) VSUMM1. $CRU_{acerto}$ = 0,75, $CRU_{erro}$ = 0,30.



(e) VSUMM2. $CRU_{acerto}$ = 0,75, $CRU_{erro}$ = 0,30.

Figura A.70: Resumos gerados pelas diferentes abordagens para o vídeo *Drift Ice as a Geologic Agent, segment 10 (v70).*