Sumarização automática multivídeo baseada em estratégias humanas

Tamires Tessarolli de Souza Barbieri

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Tamires Tessarolli de Souza Barbieri

Sumarização automática multivídeo baseada em estratégias humanas

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

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Rudinei Goularte

USP – São Carlos
Setembro de 2021

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

B236s Barbieri, Tamires Tessarolli de Souza
Sumarização automática multivídeo baseada em
estratégias humanas / Tamires Tessarolli de Souza
Barbieri; orientador Rudinei Goularte. -- São
Carlos, 2021.
138 p.

Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2021.

1. Sumarização multivídeo. 2. Critérios para
seleção de conteúdo. 3. Estratégias humanas. 4. Vídeo
digital. 5. Multimídia. I. Goularte, Rudinei,
orient. II. Título.

Tamires Tessorolli de Souza Barbieri

**Automatic multi-video summarization based on human
strategies**

Thesis submitted to the Instituto de Ciências
Matemáticas e de Computação – ICMC-USP – in
accordance with the requirements of the Computer
and Mathematical Sciences Graduate Program, for
the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and
Computational Mathematics

Advisor: Prof. Dr. Rudinei Goularte

USP – São Carlos
September 2021

Dedico este trabalho ao meu esposo Paulo e à minha filha Júlia.

AGRADECIMENTOS

Agradeço ao meu esposo Paulo pela parceria de todos os dias, por me incentivar e auxiliar quando necessário e compartilhar comigo as conquistas durante esse processo. Agradeço também à minha pequena Júlia por encher meus dias de amor e me motivar com o sorriso mais lindo.

Agradeço aos meus pais por todo o apoio, por serem meu porto seguro, ouvindo desaba-fos, dando conselhos e estando sempre presentes. Agradeço à minha irmã pelo companheirismo, pela preocupação e pelos bons momentos divididos nesse período.

Agradeço ao meu orientador por toda a contribuição que tornou possível a conclusão deste trabalho. Obrigada pela motivação, pelas correções, pelas novas ideias e pela paciência. Obrigada também pela amizade e por sempre levar em conta a vida pessoal de seus alunos, compreendendo e respeitando as particularidades de cada momento.

Agradeço aos meus amigos pelas conversas, situações compartilhadas e pelo interesse em meu trabalho. Agradeço também aos colegas do grupo de pesquisa pela troca de ideias e sugestões que ajudaram a aprimorar os resultados.

Agradeço ao Instituto de Ciências Matemáticas e de Computação (ICMC) por fornecer a estrutura necessária para a realização deste doutorado. Agradeço, em especial, aos funcionários da Seção de Pós-Graduação pelas orientações e auxílio nas dúvidas.

Agradeço ao Instituto Federal de São Paulo (IFSP) pelo afastamento remunerado para participação em programa de pós-graduação *stricto sensu*, que foi muito importante para o desenvolvimento e conclusão deste trabalho.

Agradeço à CAPES pelo financiamento parcial do estudo. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

BARBIERI, T. T. S. **Sumarização automática multivídeo baseada em estratégias humanas**. 2021. 138 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Nos últimos anos, o volume de dados multimídia produzidos e disponíveis para acesso tem crescido contínua e rapidamente. Esse contexto agravou o problema da sobrecarga de informação: encontrar conteúdo de interesse em meio à grande quantidade de opções disponíveis, tornando essenciais sistemas que possibilitem acesso eficiente. A sumarização de vídeo é uma área de pesquisa que busca lidar com esse problema, fornecendo uma versão compacta e informativa do conteúdo. Em particular, os sistemas multimídia atuais disponibilizam diversos vídeos relacionados ao mesmo assunto, contendo informações complementares. Esse fato ressalta a importância da sumarização multivídeo para lidar com o interesse do usuário em se informar sobre determinado assunto a partir de um conjunto de vídeos que o abordam, sem a necessidade de assistir a todos eles. Entretanto, a análise da literatura mostra que estratégias humanas não são consideradas na definição dos critérios utilizados para selecionar automaticamente os segmentos de vídeo que irão compor os sumários e o foco das técnicas tem sido a identificação de informações que se repetem em diferentes vídeos. Assim, esta tese tem o objetivo de investigar se critérios para seleção de conteúdo derivados de estratégias humanas são capazes de produzir resultados semanticamente mais relevantes na visão do usuário. A pesquisa desenvolvida evidencia que a abordagem proposta possibilitou mapear os julgamentos de relevância dos usuários e gerar sumários multivídeo mais próximos a suas expectativas.

Palavras-chave: Sumarização multivídeo, Critérios para seleção de conteúdo, Estratégias humanas, Vídeo digital, Multimídia.

ABSTRACT

BARBIERI, T. T. S. **Automatic multi-video summarization based on human strategies.** 2021. 138 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

In the last years, the multimedia data volume produced and available for access has increased continuous and quickly. This context has also increased the overload information problem: find content of interest in the huge amount of available options, making essential systems that allow efficient access. Video summarization is a research field that deals with this problem, providing a compact and informative version of the content. In particular, current multimedia systems make available several videos related to the same subject, having complementary information. This fact highlights the importance of multi-video summarization to deal with users' interest in being informed about a specific subject from a set of videos, without the need of watching all of them. However, the literature analysis shows that human strategies are not considered to define criteria used to automatically select video segments that will compose a summary and the focus of techniques has been the identification of common information in different videos. Therefore, this thesis aims to investigate if criteria for content selection derived from human strategies are able to produce results semantically more relevant in users' view. The developed research evidences that the proposed approach made it possible to map users' relevance judgments and generate multi-video summaries closer to their expectations.

Keywords: Multi-video summarization, Content selection criteria, Human strategies, Digital video, Multimedia.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de amostragem espacial.	26
Figura 2 – Resoluções de um vídeo digital.	27
Figura 3 – Exemplo de amostragem temporal entrelaçada.	28
Figura 4 – Estrutura do fluxo do vídeo digital.	30
Figura 5 – Exemplos de quadros-chave selecionados com o método KS-SIFT.	33
Figura 6 – Ilustração da etapa “Detecção de extremos escala-espço” do descritor SIFT.	39
Figura 7 – Exemplo de sumarização temporal.	43
Figura 8 – Exemplo de sumarização espacial.	43
Figura 9 – Classificações de sumários de vídeo.	44
Figura 10 – Exemplo de resultado da abordagem de avaliação subjetiva “Matriz de preferências”.	48
Figura 11 – Exemplo de resultado da abordagem de avaliação subjetiva baseada em notas.	49
Figura 12 – Exemplo de resultado da abordagem de avaliação subjetiva baseada em notas. Apresenta-se a média das notas atribuídas, por usuário, para os resultados de cada técnica.	50
Figura 13 – Visão geral do estudo realizado visando revelar estratégias humanas para sumarização multivídeo.	65
Figura 14 – Dados demográficos dos participantes do experimento.	67
Figura 15 – Exemplos do critério de inclusão conteúdo visual.	70
Figura 16 – Visão geral da HS-MVideoSumm, técnica de sumarização multivídeo baseada em estratégias humanas.	77
Figura 17 – Exemplo de segmento processado pela técnica HS-MVideoSumm.	78
Figura 18 – Exemplos de quadros pertencentes à introdução e segmento subsequente.	79
Figura 19 – Visão geral do módulo para detecção de segmentos subjetivos.	80
Figura 20 – Exemplos de quadros com faces em destaque.	81
Figura 21 – Critérios para classificar um segmento textual como subjetivo: opinião de jornalista ou entrevista.	83
Figura 22 – Visão geral do módulo para detecção de segmentos redundantes.	83
Figura 23 – Exemplo de união de pares de segmentos redundantes identificados pela técnica OOS. Na imagem denota-se V<núm.> - S<núm.>, representando o vídeo e seu respectivo segmento.	85
Figura 24 – Visão geral do módulo para seleção de segmentos.	85

Figura 25 – Exemplo de inserção de segmentos no sumário respeitando a ordem cronológica. Na imagem denota-se V<núm.> - S<núm.>, representando o vídeo e seu respectivo segmento.	87
Figura 26 – Exemplo de arquivo JSON com a segmentação de um vídeo em subtópicos. Contém os campos: 1) <i>content</i> (texto do segmento); 2) <i>begin</i> (tempo de início do segmento); 3) <i>end</i> (tempo final do segmento).	94
Figura 27 – <i>Ground truth</i> desenvolvido para avaliação do módulo “[2] Detecção de subjetividade”. O trecho apresentado contém a classificação de alguns segmentos do vídeo do telejornal SBT Brasil, tópico “Relação comercial entre Google e Huawei”.	96
Figura 28 – Preferências dos usuários quanto aos sumários multivídeo.	101
Figura 29 – Visão geral do método Video-MMR e parâmetros utilizados na implementação.	103
Figura 30 – Visão geral do método RSumm aplicado ao domínio de vídeos.	104
Figura 31 – Dados demográficos dos participantes do experimento de avaliação da técnica HS-MVideoSumm.	105
Figura 32 – Preferência dos usuários: por participante, quantas vezes (em porcentagem) preferiu o resultado de cada técnica de sumarização.	106
Figura 33 – Preferência dos usuários: por participante, média das notas que atribuiu aos resultados de cada técnica de sumarização.	107

LISTA DE TABELAS

Tabela 1 – Definições de verdadeiros positivos e negativos, falsos positivos e negativos.	46
Tabela 2 – Número de trabalhos excluídos por cada critério durante as etapas de seleção preliminar e seleção final do mapeamento sistemático.	52
Tabela 3 – Critérios para seleção de conteúdo identificados em trabalhos relacionados a sumarização multivídeo.	53
Tabela 4 – Comparação entre os trabalhos relacionados quanto aos critérios para seleção de conteúdo.	54
Tabela 5 – Principais características das bases de vídeo utilizadas para avaliação em trabalhos relacionados a sumarização multivídeo.	60
Tabela 6 – Descrição da base de vídeos utilizada no experimento.	64
Tabela 7 – Estratégias humanas para sumarização multivídeo identificadas no estudo com usuários.	68
Tabela 8 – Porcentagem de participantes que aplicaram cada critério. Dados obtidos pela: 1) observação dos usuários durante o experimento e suas respostas ao questionário pós-sessão e 2) investigação dos sumários produzidos.	72
Tabela 9 – Comparação entre os critérios revelados pelo estudo com usuários e os critérios identificados em trabalhos relacionados.	74
Tabela 10 – Estratégias humanas para sumarização multivídeo identificadas no estudo com usuários. Os critérios destacados em negrito foram incorporados na técnica HS-MVideoSumm.	89
Tabela 11 – Descrição da base de vídeos utilizada nos experimentos de avaliação.	92
Tabela 12 – Tempo total e número de segmentos de cada conjunto de vídeos.	94
Tabela 13 – <i>Ground truth</i> desenvolvido para avaliação do módulo “[1] Seleção da introdução”. O trecho apresentado contém as informações para o tópico “Incêndio na Catedral de Notre-Dame”.	95
Tabela 14 – Avaliação da técnica HS-MVideoSumm quanto à detecção de subjetividade. São apresentadas as métricas precisão (P), abrangência (A) e F1 calculadas sobre: 1) todos os segmentos de cada conjunto e 2) entrada do módulo “[2] Detecção de subjetividade” da técnica HS-MVideoSumm.	97
Tabela 15 – <i>Ground truth</i> desenvolvido para avaliação do módulo “[3] Detecção de segmentos redundantes”. O trecho apresentado contém os assuntos abordados em mais de um vídeo para o tópico “Relação comercial entre Google e Huawei”.	98

Tabela 16 – Avaliação da técnica HS-MVideoSumm quanto à detecção de redundância considerando-se todos os segmentos de cada conjunto de vídeos.	98
Tabela 17 – Tamanho (em minutos) dos conjuntos de vídeos e dos sumários produzidos por cada técnica avaliada.	104
Tabela 18 – Média e desvio padrão das notas que os usuários atribuíram aos sumários gerados por cada técnica para os tópicos da base de vídeos.	105

LISTA DE ABREVIATURAS E SIGLAS

HS-MVideoSumm	<i>Human Strategies based Multi-Video Summarization</i>
ASR	<i>Automatic Speech Recognition</i>
C3D	<i>Convolutional 3D</i>
CCD	<i>Charge Coupled Device</i>
CNN	<i>Convolutional Neural Network</i>
CODEC	<i>enCOder/DECoder</i>
DFT	<i>Discrete Fourier Transform</i>
DoG	<i>Difference of Gaussian</i>
fps	<i>frames per second</i>
HDS	<i>Hypergraph-based Dominant Set</i>
HOG	<i>Histogram of Oriented Gradients</i>
HSV	<i>Hue, Saturation and Value</i>
ISO	<i>International Standards Organization</i>
ITU	<i>International Telecommunications Union</i>
KS-SIFT	<i>Keyframe Set - Scale Invariant Feature Transform</i>
LBP	<i>Local Binary Pattern</i>
LIA ASR	<i>Laboratoire d'Informatique d'Avignon, France - Automatic Speech Recognition</i>
MFCC	<i>Mel-Frequency Cepstral Coefficients</i>
NLTK	<i>Natural Language Toolkit</i>
OOS	<i>One to One Symmetric matching</i>
OpenCV	<i>Open Source Computer Vision Library</i>
RGB	<i>Red, Green and Blue</i>
SIFT	<i>Scale Invariant Feature Transform</i>
SILTP	<i>Scale-Invariant Local Ternary Pattern</i>
ZCR	<i>Zero Crossing Rate</i>

SUMÁRIO

1	INTRODUÇÃO	21
1.1	Contextualização e motivação	21
1.2	Definição do problema	22
1.3	Lacunas da área de pesquisa	23
1.4	Objetivo da tese	23
1.5	Organização do texto	24
2	CONCEITOS RELACIONADOS	25
2.1	Vídeo digital	25
2.1.1	<i>Propriedades</i>	26
2.1.2	<i>Segmentação de vídeo</i>	30
2.1.3	<i>Extração de quadros-chave</i>	31
2.2	Extração de características	32
2.2.1	<i>Características textuais</i>	35
2.2.2	<i>Características visuais</i>	36
2.2.3	<i>Medidas de similaridade</i>	40
2.3	Sumarização de vídeo	42
2.3.1	<i>Processo de sumarização</i>	42
2.3.2	<i>Sumários de vídeo</i>	43
2.4	Avaliação de sumários de vídeo	45
2.4.1	<i>Avaliação objetiva</i>	46
2.4.2	<i>Avaliação subjetiva</i>	48
2.5	Considerações finais	50
3	TRABALHOS RELACIONADOS	51
3.1	Protocolo para o mapeamento sistemático	51
3.2	Resultado do mapeamento sistemático	52
3.3	Análise dos trabalhos relacionados	53
3.3.1	<i>Critérios para seleção de conteúdo</i>	53
3.3.2	<i>Bases de vídeo</i>	59
3.4	Considerações finais	60
4	CRITÉRIOS PARA SELEÇÃO DE CONTEÚDO: UM ESTUDO COM USUÁRIOS	63

4.1	Descrição do estudo	63
4.2	Análise qualitativa	67
4.3	Análise quantitativa	72
4.4	Considerações finais	75
5	TÉCNICA DE SUMARIZAÇÃO MULTIVÍDEO HS-MVIDEOSUMM	77
5.1	Seleção da introdução	78
5.2	Detecção de subjetividade	80
5.3	Detecção de segmentos redundantes	83
5.4	Seleção de segmentos	85
5.5	Produção do sumário	87
5.6	Considerações finais	88
6	AVALIAÇÃO	91
6.1	Base de vídeos	91
6.2	Avaliação objetiva	94
6.2.1	<i>Detecção da introdução</i>	95
6.2.2	<i>Detecção de subjetividade</i>	96
6.2.3	<i>Detecção de segmentos redundantes</i>	97
6.3	Avaliação com usuários	99
6.3.1	<i>Detecção de subjetividade</i>	99
6.3.2	<i>Comparação entre HS-MVideoSumm e trabalhos relacionados</i> . . .	102
6.4	Considerações finais	107
7	CONCLUSÃO	109
7.1	Contribuições	109
7.2	Limitações e trabalhos futuros	111
	REFERÊNCIAS	113
	GLOSSÁRIO	121
	APÊNDICE A QUESTIONÁRIOS - ESTUDO COM USUÁRIOS . . .	123
A.1	Termo de Consentimento Livre e Esclarecido	123
A.2	Questionário pré-sessão	125
A.3	Questionário pós-sessão	126
	APÊNDICE B FORMULÁRIOS - AVALIAÇÃO COM USUÁRIOS . .	127
B.1	Detecção de subjetividade	127
B.2	Comparação com trabalhos relacionados	133

INTRODUÇÃO

1.1 Contextualização e motivação

Nos dias atuais, os vídeos ocupam uma posição de destaque na vida cotidiana das pessoas, que os consomem com diferentes propósitos, como entretenimento, busca por informações e aperfeiçoamento profissional. Na área de entretenimento, ressaltam-se os serviços de transmissão de vídeo sob demanda. Um exemplo é a Netflix, que possui 193 milhões de assinantes em mais de 190 países, assistindo a séries de TV, documentários e filmes em diferentes dispositivos ([NETFLIX, 2020](#)). Em 2017, as pessoas ao redor do mundo assistiram a um bilhão de horas por semana nessa plataforma ([NETFLIX, 2017](#)). Outro exemplo é a Amazon Prime Video que, até janeiro de 2020, contava com mais de 150 milhões de usuários. O serviço está disponível em mais de 200 países e possui no catálogo em torno de 75 milhões de vídeos ([MARKET.US, 2020](#)).

Esse cenário também é uma realidade no Brasil. Em setembro de 2019, o YouTube revelou dados da sexta edição da pesquisa *Video Viewers* ([VIEIRA, 2019](#)). Segundo essa análise, os brasileiros já consomem diariamente mais vídeos *on-line* do que em canais de televisão. Entre os entrevistados, 95% afirmam que assistem a materiais *on-line*. De 2014 a 2019, o consumo de vídeos na web cresceu 165% e a maioria das pessoas está buscando por conteúdos que não são oferecidos pela televisão.

Entretanto, o crescimento no volume de dados multimídia produzidos e disponíveis para acesso aumenta o problema da sobrecarga de informação: as pessoas precisam encontrar informações de interesse entre as centenas ou milhares de opções disponíveis ([IDE et al., 2017](#); [KANNAN et al., 2013](#); [ZHANG et al., 2013](#); [TOFFLER, 1984](#)). Esse pode ser um processo tedioso e que demanda muito tempo. Por isso, são necessários sistemas eficientes e eficazes de acesso a conteúdos. A sumarização automática é uma área de pesquisa que busca auxiliar a diminuir esse problema ([JI et al., 2018](#); [YOSHITAKA; SAWADA, 2012](#)). Ela consiste em obter

uma alternativa concisa para o conteúdo que seja semanticamente relevante para o usuário. No caso de vídeos, se produz uma versão reduzida do vídeo original que preserve as informações mais significativas. Portanto, os sumários ajudam o usuário a encontrar e entender o conteúdo de um vídeo em um período de tempo menor. Técnicas de sumarização são relevantes para diversos tipos de aplicações, como programas de TV, filmes, vídeos de vigilância e sistemas de aprendizado eletrônico (CHEN; DE VLEESCHOUWER; CAVALLARO, 2014).

Tradicionalmente as pesquisas na área de sumarização utilizam um único vídeo como fonte para um sumário (LI *et al.*, 2021; GYGLI *et al.*, 2014; POTAPOV *et al.*, 2014; ALMEIDA; LEITE; TORRES, 2013). Contudo, em domínios relevantes como vigilância, notícias e educação, é comum encontrar diversos vídeos referentes ao mesmo tópico. Assistir a todos eles pode ser uma tarefa inviável em termos de tempo e, em muitos casos, repetitiva devido à sobreposição de conteúdos. Por outro lado, tais vídeos englobam diferentes modos de retratar um assunto, contendo informações complementares, o que pode ser explorado para enriquecer os sumários produzidos. Tais características tornam a sumarização multivídeo uma necessidade e foco de pesquisas (JI *et al.*, 2020; WU; ZHONG; LIU, 2020; JI *et al.*, 2019; IDE *et al.*, 2017; PANDA; MITHUN; ROY-CHOWDHURY, 2017; LI; MERIALDO, 2016), cujo objetivo é produzir um único sumário a partir de um conjunto de vídeos relacionados. Esse é um processo complexo (LI; MERIALDO, 2016), no qual existem dois desafios principais (SHAO *et al.*, 2010): 1) a diversidade de tópicos e sobreposição de conteúdos em diferentes vídeos; 2) a natureza multimodal dos vídeos, que exige que as técnicas integrem características de diversas fontes (texto, áudio, imagem). Além disso, conforme será discutido no Capítulo 3, a sumarização multivídeo ainda é pouco investigada, em comparação a outras tarefas como a segmentação em tomadas, apresentando diversas lacunas de pesquisa a serem exploradas.

1.2 Definição do problema

Dado, como entrada, um conjunto de vídeos relacionados a um mesmo tópico ou evento de interesse, a sumarização multivídeo consiste em selecionar, entre os segmentos¹ de todos os vídeos, aqueles que irão compor o sumário.

A seleção de segmentos, também chamada seleção de conteúdo, deve ser baseada em critérios que permitam automaticamente determinar a relevância de cada segmento e, deste modo, escolher os que conduzam a um sumário compacto e, ao mesmo tempo, informativo.

¹ Um segmento de vídeo é definido como um conjunto de quadros consecutivos que são homogêneos de acordo com algum critério estabelecido (DIMITROVA *et al.*, 2009).

1.3 Lacunas da área de pesquisa

Da análise da literatura (detalhada no Capítulo 3), observa-se que uma das lacunas da área de pesquisa está relacionada a um passo essencial nas técnicas de sumarização multivídeo: a definição de critérios para seleção de conteúdo capazes de indicar a relevância de um segmento de vídeo para fins de inclusão no sumário. Trabalhos relacionados são, em sua maioria, baseados em redundância e diversidade (WU; ZHONG; LIU, 2020; JI *et al.*, 2019; KANEHIRA *et al.*, 2018; LI; MERIALDO, 2016), visando identificar informações que se repetem em diferentes vídeos. Porém, sumários produzidos por tais critérios podem não satisfazer as expectativas do usuário, por não mapearem suas preferências e necessidades e, conseqüentemente, não incorporarem o nível de semântica esperado.

Por isso, uma abordagem alternativa, inexplorada no contexto de sumarização multivídeo, mas utilizada com sucesso na área de sumarização multidocumento (CAMARGO; DI-FELIPPO; PARDO, 2015; AGOSTINI; LÓPEZ CONDORI; PARDO, 2014), é a obtenção de critérios para seleção de conteúdo a partir de estratégias humanas. Isso porque as pessoas são capazes de compreender a semântica e as correlações entre diferentes segmentos e utilizar esse conhecimento no processo de produção de sumários. Logo, derivar os critérios de seleção de conteúdo de estratégias humanas tem potencial para aprimorar os resultados da sumarização multivídeo, tornando-os mais próximos à percepção do usuário e mais significativos semanticamente.

1.4 Objetivo da tese

O objetivo desta tese é investigar se critérios para seleção de conteúdo derivados de estratégias humanas são capazes de produzir sumários semanticamente mais relevantes na visão do usuário final. Para isso, propõe-se desenvolver e avaliar uma técnica de sumarização multivídeo baseada em tais critérios e aplicada ao domínio de vídeos de notícias.

Destacam-se duas etapas principais do trabalho: 1) estudo com usuários; 2) técnica de sumarização multivídeo.

Estudo com usuários. Desenvolvimento de um estudo com usuários finais visando compreender como realizam o processo de sumarização de um conjunto de vídeos relacionados para derivação de critérios de seleção de conteúdo.

Técnica de sumarização multivídeo. Desenvolvimento e avaliação de uma técnica automática de sumarização multivídeo que incorpore os critérios de seleção de conteúdo obtidos.

1.5 Organização do texto

O restante do texto está organizado da seguinte maneira: no Capítulo 2, são abordados conceitos fundamentais no contexto do trabalho; no Capítulo 3, apresenta-se uma análise de trabalhos relacionados a sumarização multivídeo reportados na literatura, concentrando-se nos critérios para seleção de conteúdo utilizados; no Capítulo 4, discute-se o estudo com usuários desenvolvido com o objetivo de compreender as estratégias por eles empregadas no processo de sumarização de um conjunto de vídeos relacionados; no Capítulo 5, detalha-se a técnica de sumarização multivídeo desenvolvida, a qual incorpora os critérios para seleção de conteúdo derivados de estratégias humanas; no Capítulo 6, a técnica é avaliada no domínio de vídeos de notícias, por meio de: 1) medidas objetivas de eficácia e 2) experimentos com usuários visando comprovar se os sumários produzidos estão alinhados às suas expectativas; no Capítulo 7, destacam-se as contribuições do trabalho desenvolvido, bem como suas limitações.

CONCEITOS RELACIONADOS

Neste capítulo, são discutidos conceitos relevantes no contexto deste trabalho. Na Seção 2.1, apresentam-se propriedades do vídeo digital, sua segmentação em diferentes níveis, bem como métodos para seleção de quadros-chave. Na Seção 2.2, são discutidos extratores de características para representação do conteúdo de vídeo, em especial, para descrição de informações textuais e visuais, além de medidas de similaridade que viabilizam a comparação entre as representações obtidas. Na Seção 2.3, são aprofundados conceitos relacionados à sumarização de vídeo, trazendo possíveis classificações de sumários. Na Seção 2.4, discutem-se métricas para avaliação objetiva dos resultados da sumarização multivídeo e abordagens envolvendo usuários. Por fim, a Seção 2.5 traz algumas considerações sobre os assuntos explorados no capítulo.

2.1 Vídeo digital

Nas últimas décadas, tem-se vivenciado uma revolução nas mídias digitais. O vídeo digital, em particular, aproximou as indústrias de transmissão, cinema, computador e comunicação, tornando telefone, TV a cabo e Internet serviços concorrentes. Atualmente, um único dispositivo pode ser usado como computador pessoal, TV em alta definição e meio de comunicação (TEKALP, 2015), facilitando a produção de vídeo digital.

Segundo Compesi (2015), o processo de produção de um vídeo pode ser dividido em três estágios: pré-produção, produção e pós-produção. A etapa de pré-produção é crucial por estar relacionada ao planejamento, incluindo definição do assunto, preparação de *scripts* e estudos de viabilidade. A fase de produção envolve a gravação do vídeo em si. A pós-produção, por sua vez, está relacionada a aspectos de edição e distribuição.

É importante ressaltar que a produção de um vídeo não é mais uma atividade cara e trabalhosa, acessível apenas para emissoras de televisão licenciadas. Como resultado da portabilidade e disponibilidade de novas tecnologias, expandiu-se o interesse pelo vídeo. Cada vez mais,

eles estão sendo utilizados para expressão pessoal, produções independentes e em instituições educacionais e corporações. Os atuais dispositivos portáteis para gravação de vídeo têm ainda a vantagem de produzir arquivos que estão imediatamente disponíveis para visualização. Exemplos comuns de utilização incluem: registro de eventos familiares, como aniversários, festas e casamentos ou criação de mensagens para enviar a amigos ou parentes.

A área de produção de vídeos combina o entendimento de fatores estéticos e técnicos. Fatores estéticos estão relacionados às variáveis de produção e os meios pelos quais podem ser manipuladas para afetar a resposta da audiência ao conteúdo do vídeo. Fatores técnicos, por sua vez, abordam o entendimento do sinal de vídeo e como pode ser controlado (COMPELI, 2015), sendo essenciais para qualquer aplicação que necessite processá-lo. Logo, as principais propriedades de um vídeo digital são discutidas na Seção 2.1.1.

2.1.1 Propriedades

Um vídeo digital pode ser definido como uma sequência de imagens frequentemente acompanhadas de áudio e legenda (texto) (BLANKEN *et al.*, 2007) ou uma representação de uma cena visual natural ou do mundo real, amostrada espacial e temporalmente (RICHARDSON, 2010).

Concentrando-se na aquisição de aspectos visuais, para se obter uma amostra de uma imagem 2D, utiliza-se uma câmera para focar uma projeção da cena em um sensor, como o dispositivo de carga acoplada, do inglês *Charge Coupled Device* (CCD). Para captura de imagens coloridas, cada componente de cor é separadamente filtrada e projetada no sensor.

Figura 1 – Exemplo de amostragem espacial.

(a) Grade de amostragem.

(b) Imagem amostrada com a grade em 1a.

Fonte: Richardson (2010).

A saída de um CCD é um sinal de vídeo analógico, isto é, um sinal elétrico que representa a imagem. A amostragem desse sinal em um instante de tempo produz uma imagem ou quadro que possui valores definidos por um conjunto de pontos. O formato mais comum de

amostragem espacial é um retângulo com pontos posicionados em uma grade quadrada ou retangular. A amostragem ocorre em cada intersecção da grade e a imagem pode ser reconstruída pela representação de cada unidade como um elemento quadrado: *pixel*. O número de pontos amostrados influencia a qualidade visual. Escolher uma grade grosseira gera uma imagem de baixa resolução, como ilustrado na Figura 1. Nela pode-se verificar que a grade utilizada possui poucos pontos de amostragem, consequentemente, o quadro resultante contém um baixo número de *pixels* (RICHARDSON, 2010).

Logo, a **resolução** horizontal e vertical de um vídeo digital está relacionada à densidade de amostragem. O número de *pixels* por linha e o de linhas por quadro é usado para classificar a definição do vídeo como padrão, alta ou ultra alta, como pode ser visto na Figura 2. Vale ressaltar que, em vídeos de baixa resolução, surgem artefatos (do inglês, *aliasing*), como bordas irregulares, cuja visibilidade varia com o tamanho da tela e distância de visualização (TEKALP, 2015).

Figura 2 – Resoluções de um vídeo digital.

Fonte: Tekalp (2015).

A sensação de movimento em um vídeo é produzida pela obtenção de amostras do sinal em intervalos de tempo periódicos. A reprodução de uma série de quadros gera a aparência de movimento. Uma taxa de **amostragem temporal** mais alta, também chamada **taxa de quadros** (do inglês, *frames per second* (fps)), proporciona apresentação mais suave, porém requer que mais amostras sejam capturadas e armazenadas. Taxas abaixo de 10 quadros por segundo podem ser usadas para comunicação com taxa de *bits* limitada, pois a quantidade de dados é relativamente pequena. Entretanto, nesse caso, a exibição é irregular e não natural. Taxas entre 10 e 20 quadros por segundo também são aplicáveis ao contexto de comunicação; nesse intervalo a imagem é mais suave, mas fluxos irregulares podem ser visíveis em trechos com alta movimentação. A norma padrão para televisão é de 25 ou 30 quadros por segundo. Por fim, tem-se que amostragens temporais de 50 ou 60 quadros por segundo resultam em uma aparência de movimentação muito suave, porém necessitam de um volume de dados alto.

Um sinal de vídeo pode ser amostrado como uma série de quadros completos (**amostragem progressiva**) ou uma sequência de campos entrelaçados (**amostragem entrelaçada**). Em um

trecho de vídeo entrelaçado, metade dos dados em um quadro, isto é, um campo, é obtida a cada intervalo de amostragem. Um campo pode ser composto pelas linhas ímpares ou pares da imagem, como exemplificado na Figura 3. Esse método tem como vantagem a possibilidade de enviar duas vezes mais campos por segundo do que o número de quadros na equivalente sequência progressiva, mantendo-se a quantidade de dados, o que proporciona uma aparência de movimento mais suave (RICHARDSON, 2010).

Figura 3 – Exemplo de amostragem temporal entrelaçada.

Fonte: Richardson (2010).

Muitas aplicações de vídeo digital dependem da exibição de cores, por isso são necessários mecanismos capazes de capturar e representar essa informação. Uma imagem monocromática requer apenas um número para indicar a luminância de cada amostra espacial, enquanto imagens coloridas necessitam de, pelo menos, três valores por *pixel*. O método escolhido para representar luminância e crominância é chamado, segundo Richardson (2010), **espaço de cor**.

No espaço *Red, Green and Blue* (RGB), um *pixel* é descrito por três números que denotam a proporção das cores primárias vermelho, verde e azul. A captura de uma imagem RGB envolve filtrar cada componente de cor de uma cena em um sensor individual. Na exibição, em uma distância de visualização apropriada, as componentes se unem proporcionando a aparência da cor real.

O sistema visual humano é menos sensível a cor do que a luminância (RICHARDSON, 2010). No espaço RGB, as três componentes são consideradas igualmente importantes e, em geral, armazenadas com a mesma resolução. Entretanto, é possível representar uma imagem colorida de modo mais eficiente separando-se a informação de luminância e descrevendo-a com uma resolução maior que a de cor. O espaço YCrCb é comumente utilizado para essa finalidade. O Y representa a luminância e Cr, Cb e Cg caracterizam, para cada pixel, a diferença entre a intensidade das cores vermelho, azul e verde, respectivamente, e o valor de luminância. Como a soma $Cr + Cb + Cg$ é uma constante, apenas duas das componentes de crominância precisam ser armazenadas e transmitidas, no caso Cr e Cb. Além disso, devido às características do sistema

visual humano, Cr e Cb podem ser representadas com menos amostras, reduzindo a quantidade de dados sem um impacto óbvio na qualidade visual.

Assim, é possível converter uma imagem RGB para YCrCb após a captura, de maneira a diminuir os recursos necessários para armazenamento e transmissão. Também pode-se convertê-la novamente ao espaço RGB antes da apresentação. As equações de conversão RGB-YCrCb são padronizadas pela *International Telecommunications Union* (ITU) e amplamente utilizadas em sistemas de codificação de vídeo. Richardson (2010) apresenta detalhadamente tais equações.

Logo, como evidenciado nas discussões anteriores, a representação de sinais de vídeo digital requer uma grande quantidade de dados. Por isso, necessitam-se de meios de compressão para que tais dados possam ser utilizados por diferentes tipos de aplicação. **Compressão ou codificação** de vídeo é o processo de convertê-lo em um formato apropriado para transmissão ou armazenamento, reduzindo a quantidade de *bits* necessária para representar a informação. Envolve um par complementar de sistemas: compressor/codificador e descompressor/decodificador. O primeiro converte os dados de origem em uma forma comprimida, ocupando um número menor de *bits*, enquanto o segundo recupera uma representação do vídeo original. O par codificador/decodificador é conhecido como *enCOder/DECOder* (CODEC) (RICHARDSON, 2010).

A compressão pode ser classificada como **sem perdas**, no caso em que o mesmo dado é recuperado após a descompressão ou **com perdas**, quando se obtém uma aproximação do sinal original (AKRAMULLAH, 2014). Na situação sem perdas, removem-se redundâncias estatísticas presentes nos dados, porém obtém-se uma taxa de compressão moderada. Por exemplo, para imagens, o máximo que se consegue alcançar com codificadores padrão é uma redução de 3 a 4 vezes. Assim, o caso com perdas é necessário para se atingir taxas de compressão mais altas. Esse tipo de sistema baseia-se na remoção de redundâncias subjetivas, isto é, elementos da imagem ou sequência de vídeo que podem ser excluídos sem afetar significativamente a percepção do usuário a respeito da qualidade visual.

A maioria dos métodos de codificação de vídeo explora redundâncias temporais e espaciais. No domínio temporal, em geral, há uma alta correlação ou similaridade entre quadros capturados em tempos próximos, especialmente se a taxa de quadros for alta. No domínio espacial, por sua vez, verifica-se que os valores de *pixels* vizinhos são frequentemente similares (RICHARDSON, 2010).

A ubiquidade dos vídeos tornou-se possível devido ao estabelecimento de diretrizes para representação. Padrões internacionais de codificação de vídeo são definidos por comitês de organizações como *International Standards Organization* (ISO) e ITU. O objetivo da padronização é atender a diversas necessidades da indústria. Por exemplo, a especificação de formatos é essencial para a transmissão de vídeos entre diferentes produtos e aplicações; devido ao volume de dados, são necessários também modelos de compressão. Para determinar esses padrões, os grupos de especialistas revisaram uma variedade de técnicas de compressão, estruturas de dados

e algoritmos, concordando sobre algumas tecnologias. As atividades de padronização evoluíram com o rápido avanço nas áreas de televisão, filmes, computadores, comunicação e processamento de sinais. Assim, podem ser citados como exemplos de padrões de codificação de vídeo: MPEG-1, H.263, MPEG-2, MPEG-4 Part 2, AVC/H.264, HEVC/H.265 e H.266 ([AKRAMULLAH, 2014](#)).

2.1.2 Segmentação de vídeo

Devido ao alto volume de dados, muitas aplicações, incluindo sumarização, se utilizam da divisão do vídeo em unidades menores e mais facilmente gerenciáveis. Esse processo, conhecido como segmentação temporal, consiste em particionar uma sequência de vídeo em conjuntos de quadros consecutivos que são homogêneos de acordo com algum critério definido ([DIMITROVA et al., 2009](#)). A segmentação do conteúdo pode ocorrer em diversos níveis. Em geral, a estrutura hierárquica do fluxo de vídeo é composta de quatro segmentos distintos ([ZHAO; YANG; FENG, 2001](#)), conforme explicado a seguir e ilustrado na Figura 4.

- Quadro, definido como uma única imagem estática ([RICHARDSON, 2002](#)).
- Tomada, definida como um conjunto de quadros, representando uma ação contínua capturada por uma única câmera ([SMEATON, 2007](#)).
- Cena, comumente definida como um grupo de tomadas semanticamente relacionadas ([RUI; HUANG; MEHROTRA, 1999](#); [ZHAI; SHAH, 2006](#)).
- Vídeo completo.

Figura 4 – Estrutura do fluxo do vídeo digital.

Fonte: [Coimbra \(2011\)](#).

Entretanto, destaca-se que o tamanho do segmento de vídeo é usualmente determinado pela aplicação. Além da estrutura hierárquica, pode ser utilizado como critério para a divisão: o tempo (por exemplo, 1 ou 2 segundos ([LI; MERIALDO, 2016](#))) ou outra informação relativa ao contexto. No caso deste trabalho, considera-se que um segmento é um trecho de vídeo, composto por imagens e respectivo conteúdo de áudio e texto. Mais especificamente, no contexto da técnica de sumarização multivídeo desenvolvida, seu tamanho é definido em função da semântica expressa nas informações textuais, conforme será discutido no Capítulo 5.

2.1.3 Extração de quadros-chave

Após o processo de segmentação temporal, a análise dos aspectos visuais de um vídeo geralmente envolve a extração de quadros-chave, que tem como objetivo reduzir a quantidade de dados a serem processados. Um quadro-chave é definido, segundo [Hua e Wang \(2009\)](#), como o quadro que melhor representa o conteúdo visual de um segmento de vídeo. Dependendo da complexidade das informações presentes, se estabelece a quantidade de quadros-chave a ser selecionada, podendo ser apenas um ou vários. Essa representação pode, então, ser usada como entrada para a manipulação dos dados de um vídeo, por exemplo, para tarefas de indexação ou recuperação.

Os métodos amplamente adotados para a extração de quadros-chave podem ser categorizados, de acordo com [Hua e Wang \(2009\)](#), em: 1) baseados em análise; 2) baseados em agrupamento, conforme explicado a seguir.

Método baseado em análise. Consiste na extração de quadros-chave pela análise do conteúdo do vídeo, como a qualidade ou a atratividade dos quadros.

Método baseado em agrupamento. Um processo de agrupamento de quadros é realizado e os centróides dos grupos são estabelecidos como quadros-chave.

Neste trabalho, utilizou-se um método baseado em análise para a seleção dos quadros-chave de cada segmento de vídeo, denominado *Keyframe Set - Scale Invariant Feature Transform* (KS-SIFT) ([BARBIERI, 2014](#)). Tal método destaca-se por considerar a relação entre os quadros, comparando o conteúdo visual para determinar os que são suficientemente diferentes e representam pontos importantes do segmento. O KS-SIFT pode ser dividido em três etapas, descritas a seguir, segundo [Barbieri \(2014\)](#).

Etapa 1: Primeiro quadro-chave

Com o objetivo de que todos os segmentos sejam representados por, pelo menos, um quadro-chave, define-se que o primeiro quadro de cada segmento deve ser adicionado ao conjunto de quadros-chave.

Vale ressaltar que o método se baseia no descritor de características locais *Scale Invariant Feature Transform* (SIFT) (Seção 2.2.2) por considerar que é capaz de encontrar pontos de interesse que identificam unicamente uma imagem, favorecendo a escolha dos quadros que melhor representam a variação de conteúdo dentro de um segmento.

Etapa 2: Análise do número de pontos de interesse

Após a definição do primeiro quadro-chave, avaliam-se quais os demais quadros a serem adicionados ao conjunto (adotando-se uma taxa de amostragem de 1 quadro por segundo), de modo a contribuir com a representação. Logo, o segundo passo do método KS-SIFT

baseia-se no seguinte raciocínio: se o número de pontos de interesse do quadro sendo analisado, referenciado como quadro atual, é suficientemente diferente do número de pontos de interesse de todos os quadros já selecionados, assume-se que o quadro atual é diferente dos demais já escolhidos e adiciona-se o mesmo ao conjunto de quadros-chave.

Mais especificamente, verifica-se se o módulo da diferença entre o número de pontos de interesse do quadro atual e do quadro-chave é maior ou igual a 60% do número de pontos do quadro-chave. É importante ressaltar que esse critério deve ser satisfeito entre o quadro atual e todos os quadros já pertencentes ao conjunto de quadros-chave para que ele seja inserido.

Etapa 3: Análise dos vetores de características

Se o quadro atual apresenta um número de pontos de interesse semelhante ao dos quadros já selecionados, verifica-se se esses pontos são diferentes. Isso porque, duas imagens com conteúdos distintos podem possuir um número semelhante de pontos de interesse, porém, nesse caso, eles não representarão as mesmas informações. Para fazer essa avaliação, os vetores de características que descrevem os pontos de interesse são analisados com a técnica *One to One Symmetric matching* (OOS) (Seção 2.2.3).

Mais especificamente, com a OOS, faz-se a correspondência entre os vetores de características do quadro atual e do quadro-chave e verifica-se se o número de vetores correspondentes é menor que 10% do número de vetores do quadro-chave. O raciocínio é o seguinte: se muitos vetores de características do quadro-chave estão presentes no quadro atual, então esse quadro já está representado pelo quadro-chave, ou seja, é semelhante a ele e não precisa ser incluído no conjunto de imagens selecionadas. De maneira análoga, se poucos vetores de características do quadro-chave estão presentes no quadro atual, considera-se que as imagens são diferentes e o quadro atual deve ser inserido no conjunto. Novamente, o critério deve ser satisfeito entre o quadro atual e todos os quadros já selecionados para que o mesmo seja inserido no conjunto de quadros-chave.

Na Figura 5 apresentam-se exemplos de quadros-chave selecionados pelo KS-SIFT para alguns segmentos de vídeo. Pode-se observar que o método é capaz de capturar a diversidade de conteúdo e evitar repetições, gerando um conjunto compacto e representativo.

2.2 Extração de características

Após capturados e armazenados, como discutido nas seções anteriores, os vídeos digitais estão sujeitos a busca e recuperação pelos usuários. Entretanto, a recuperação de dados multimídia é diferente da de dados estruturados. Realizar buscas em uma base relacional pode ser considerado um processo bem estabelecido: tem-se conhecimento da estrutura e usando uma linguagem (por

Figura 5 – Exemplos de quadros-chave selecionados com o método KS-SIFT.

(a) Quadros-chave extraídos para um segmento do vídeo do Jornal da Band, tópico “Apresentação do bebê da família real britânica”.

(b) Quadros-chave extraídos para um segmento do vídeo do Jornal Nacional, tópico “Incêndio na Catedral de Notre-Dame”.

Fonte: Dados da pesquisa.

exemplo, SQL) pode-se especificar a requisição. Já para tipos de dados como imagem, áudio e vídeo, o método tradicional de recuperação não é aplicável.

Por isso, empregam-se abordagens que tentam derivar características dos próprios objetos multimídia, do inglês *features*. O processo de obtê-las é denominado extração de características e geralmente é executado de maneira automática. Segundo Blanken *et al.* (2007), as *features* podem ser divididas em duas classes: baixo e alto nível.

Características de **baixo nível** são fortemente dependentes do tipo de dados (texto, áudio e imagem) e obtêm automaticamente padrões e estatísticas de um objeto multimídia, os quais são expressos em vetores contendo valores numéricos (vetores de características). Para textos, tem-se que, durante o processo de indexação, palavras como “o(a)”, “um(a)”, etc. são excluídas por não acrescentarem significado ao documento. Assim, para descrever o conteúdo, utiliza-se uma lista de palavras-chave com indicadores de frequência.

Um sinal de áudio, por sua vez, pode ser representado por uma sequência amplitude-tempo: para cada amostra, a pressão do ar é quantificada. A amplitude equivalente à pressão do ar em caso de silêncio é descrita por zero, sendo que valores de pressão mais altos que o silêncio indicam amplitudes positivas e vice-versa. Assim, algumas características de baixo nível podem ser derivadas dessa sequência, por exemplo: 1) energia média, descreve a intensidade do sinal;

2) *Zero Crossing Rate* (ZCR), obtém a frequência de mudança de sinal (positivo/negativo) na amplitude; 3) taxa de silêncio. Uma baixa taxa de silêncio indica música, enquanto uma alta variabilidade ZCR, em geral, sugere diálogo. A Transformada Discreta de Fourier (do inglês, *Discrete Fourier Transform* (DFT)) da representação amplitude-tempo também pode ser usada para extrair características. Nesse caso, são obtidas as frequências do sinal, que podem auxiliar na determinação de informações como o tipo de som (fala, música, etc.), identificação de pessoas pela fala, sons audíveis e inaudíveis, entre outras.

No caso de imagens, pode-se contar o número de *pixels* que possuem cada cor em certo intervalo (histogramas de cor). Outra possibilidade é a identificação de relações espaciais: por exemplo, quando uma área azul aparece acima de uma área amarela em uma foto de praia. Além disso, se uma imagem possui muitos pontos escuros com vizinhos claros, conclui-se que apresenta altos valores de contraste. Diversas outras características, como forma e textura, são utilizadas na representação de imagens.

Entretanto, características de baixo nível geralmente não possuem significado claro para o usuário. Por exemplo, considerando-se imagens, o que um histograma de cor realmente significa? Muito verde indica a presença de um campo de golfe ou uma floresta? Com características de **alto nível** podem ser obtidos conceitos relevantes para o usuário. Porém, torna-se evidente que há uma distância entre a representação computacional dos dados (*features* de baixo nível) e seu significado (alto nível), problema conhecido como lacuna semântica (SMEULDERS *et al.*, 2000; BLANKEN *et al.*, 2007).

Um texto, como abordado anteriormente, pode ser descrito por palavras-chave e suas frequências. Nesse caso, há um relacionamento inerente com os conceitos reconhecidos por humanos. Por exemplo, quando um documento contém palavras como “futebol” e “juiz”, tem-se indícios de seu conteúdo.

Na área de reconhecimento de fala progressos foram feitos. Muitas linguagens possuem conversores de informações de áudio em texto. Assim, assumindo-se que determinado áudio contém falas, pode-se extrair de seu conteúdo características de baixo nível e relacioná-las com palavras e sentenças.

Para dados como imagens e áudio sem falas, derivar conceitos a partir de características de baixo nível é mais complicado. Nesse caso, a identificação de significados é dependente do domínio. Por exemplo, para vídeos de jogos de futebol, um som alto vindo da platéia, combinado com um objeto redondo passando uma linha branca e um apito, indica um conceito semântico relevante: um gol (BLANKEN *et al.*, 2007).

A seguir, são discutidos em mais profundidade extratores de características textuais (Seção 2.2.1) e visuais (Seção 2.2.2) amplamente utilizados e bem estabelecidos na literatura.

2.2.1 Características textuais

Para representar informações textuais, uma possível abordagem, conhecida na literatura como modelo *Bag of Words*, consiste em gerar uma síntese quantitativa de um documento. Nela, a ordem dos termos é ignorada e retêm-se apenas dados sobre o número de ocorrências. Logo, a sentença “Maria é mais veloz que João” é considerada idêntica à “João é mais veloz que Maria”. No entanto, é intuitivo inferir que dois documentos com representações *Bag of Words* similares são também próximos em conteúdo. Nesta seção, os conceitos são apresentados segundo Manning, Raghavan e Schütze (2009).

Antes de serem representados com o modelo *Bag of Words*, os textos, em geral, passam por um pré-processamento. Algumas operações usualmente realizadas nessa etapa são descritas a seguir.

Remoção de stopwords. Algumas palavras extremamente comuns, que possuem pouco valor discriminativo, são excluídas do vocabulário. Tais palavras são chamadas *stopwords*. O uso de uma lista de *stopwords* reduz a necessidade de armazenamento e, em geral, não traz prejuízos para a caracterização. Exemplos de palavras consideradas *stopwords* incluem: um, uma, e, como, em, por, de, o, a, com.

Case-folding. Uma estratégia comum é reduzir todas as letras a sua forma minúscula. Essa operação permite mapear, por exemplo, instâncias que aparecem no início de frases com outras que ocorrem no meio do texto.

Stemming. Por razões gramaticais, os documentos utilizam diferentes formas de uma palavra. O objetivo do *stemming* é reduzir flexões e derivações para uma forma base comum. Trata-se de um processo heurístico que corta o final das palavras e inclui a remoção de afixos derivacionais.

No modelo *Bag of Words* atribui-se um peso a cada termo de um documento. A abordagem mais simples é considerar que esse peso é igual ao número de ocorrências do termo t no documento d . Esse esquema é chamado frequência do termo (do inglês, *term frequency*) e é denotado por $tf_{t,d}$.

Entretanto, a abordagem tf considera que todos os termos são igualmente relevantes enquanto, na verdade, alguns possuem menor poder discriminativo. Por exemplo, um conjunto relacionado à indústria automotiva provavelmente conterá “carro” em quase todos os textos. Por isso, é introduzido um mecanismo para atenuar o efeito de termos que ocorrem frequentemente na coleção. Para esse propósito, utiliza-se a frequência do documento (do inglês, *document frequency*), denotada por df_t e definida como o número de documentos na coleção que contém o termo t .

O fator df é usado para ponderar o peso: considerando-se N o número total de documentos em uma coleção, apresenta-se a frequência inversa do documento (do inglês, *inverse document frequency*), denotada por idf_t e definida na Equação 2.1. Assim, tem-se que o idf de um termo raro é alto, enquanto o idf de um termo frequente é baixo.

$$idf_t = \log \frac{N}{df_t} \quad (2.1)$$

Combinando-se as definições de frequência do termo e frequência inversa do documento, pode-se produzir um peso composto para cada termo. O esquema $tf-idf$ atribui a um termo t em um documento d um peso dado pela Equação 2.2.

$$tf-idf_{t,d} = tf_{t,d} \times idf_t \quad (2.2)$$

Em outras palavras, o valor $tf-idf_{t,d}$ é:

- maior quando t ocorre muitas vezes em um pequeno número de documentos;
- menor quando o termo ocorre menos vezes em um documento ou ocorre em muitos documentos.

Como resultado tem-se um vetor que captura a importância relativa dos termos. Isto é, cada documento é representado como um vetor em que cada posição corresponde a um termo no vocabulário e possui peso dado pela Equação 2.2.

2.2.2 Características visuais

Nesta seção, discutem-se histogramas e descritores locais. Tratam-se de duas abordagens diferentes para representar o conteúdo visual de um vídeo, capazes de capturar informações de cor e identificar/descrever pontos de interesse, respectivamente. Além disso, são extratores em destaque na literatura, aplicados em diversos trabalhos relacionados a sumarização multivídeo.

Histogramas

Histogramas podem ser calculados sobre quadros de vídeos em escala de cinza ou coloridos de algum espaço de cores, por meio da decomposição dos canais de cor (MARQUES FILHO; VIEIRA NETO, 1999).

O histograma em escala de cinza de uma imagem é uma função discreta formada pela contagem do número de *pixels* na imagem que possuem o mesmo valor de intensidade, representando um determinado nível de cinza. Quando essa função é normalizada para somar 1 para todos os níveis de cinza, ela pode ser tratada como uma função densidade de probabilidade que expressa a probabilidade de um certo tom de cinza ser encontrado na imagem (PETROU;

PETROU, 2010). Logo, o histograma de intensidade mostra quanto cada nível de cinza ocupa da imagem. Além disso, o contraste pode ser medido pelo intervalo entre os níveis.

Porém, além do histograma em escala de cinza, histogramas de cor também podem ser extraídos. Um histograma de cor, segundo Marques (2011), pode ser definido como uma representação gráfica da frequência de cada nível de cor em uma imagem. Entretanto, enquanto histogramas em escala de cinza possuem apenas um canal de cor com 256 níveis, histogramas de cor, por exemplo, no espaço RGB, possuem três canais, isto é, o triplo de dados. Dependendo da imagem a ser processada, a combinação de cores para compor o histograma pode resultar em um volume alto de informação.

Assim, para reduzir tal volume de dados, as imagens coloridas sofrem o processo de quantização (MARQUES, 2011): cores próximas são agrupadas em um conjunto e, dentro desse conjunto, todas as cores são igualadas. Ao número de grupos da quantização costuma-se dar o nome de *bins* (caixas). Logo, uma imagem no espaço de cor RGB com 12 *bins* na proporção 4:4:4 (isto é, 4 para vermelho, 4 para verde e 4 para azul) seria representada em apenas 64 valores, ao invés de mais de 16 milhões caso a quantização não fosse utilizada. O tamanho do histograma é obtido por meio da combinação de todos os quatro valores de vermelho, quatro de verde e quatro de azul, formando então 64 possibilidades.

O processo de quantização apresenta vantagens tanto na redução do tamanho do histograma como em velocidade de processamento de operações sobre o histograma resultante. Como desvantagem, cita-se que a divisão de um histograma em *bins* resulta na perda irreversível de informações, como pequenos detalhes da imagem e degradês.

Duas das principais operações sobre histogramas são a intersecção e a diferença absoluta (MARQUES, 2011). Na intersecção, dois histogramas de imagens diferentes sofrem a operação de intersecção para cada um de seus valores, sendo que os dados comuns são armazenados em um novo histograma. Caso se deseje um único valor, esse será a soma de todos os valores para cada uma das cores presentes no histograma da intersecção.

Já no caso da diferença, dois histogramas sofrem a operação de diferença entre seus valores. Caso se deseje um valor único, o resultado de cada diferença é somado, em módulo. Vale ressaltar que o cálculo é feito em módulo para que não ocorram negativos.

Os histogramas consistem em uma abordagem de caracterização simples, mas são amplamente empregados, principalmente por serem fáceis de computar, exigirem um baixo custo de processamento e resultarem em uma representação compacta. Além disso, a cor é a característica de baixo nível mais comumente utilizada em sistemas de recuperação de imagens baseada em conteúdo (STEHLING; NASCIMENTO; FALCÃO, 2002) e tem grande importância também no domínio de vídeos digitais. Entretanto, os histogramas não são capazes de capturar a semântica do conteúdo e, em alguns casos, imagens diferentes possuem histogramas semelhantes ou idênticos. Essas situações prejudicam a representatividade da caracterização.

Descritores Locais

Um modo diferente do anteriormente apresentado para se representar imagens é utilizar extratores de características visuais locais, que estão associados à identificação e descrição de pontos de interesse. Embora existam diversas técnicas de descrição local de imagens, todas se baseiam em três passos principais, explicados a seguir, segundo [Bay, Tuytelaars e Van Gool \(2006\)](#).

- Pontos de interesse são selecionados em alguns locais da imagem, tais como cantos ou junções-T. A propriedade mais valiosa de um detector de pontos de interesse é a repetitividade, isto é, ser confiável para encontrar os mesmos pontos de interesse em diferentes condições de visualização.
- A vizinhança de cada ponto de interesse é representada por um vetor de características. Esse descritor deve ser diferenciado e, ao mesmo tempo, robusto a ruídos, detecção de erros e deformações geométricas e fotométricas.
- Os vetores de características de diferentes imagens são comparados. A comparação é frequentemente baseada no uso de medidas de similaridade ([MANNING; RAGHAVAN; SCHÜTZE, 2009](#)), por exemplo, distância Euclidiana. A dimensão do vetor tem um impacto direto no tempo de execução deste passo.

Scale Invariant Feature Transform (SIFT)

Como exemplo de descritor visual local apresenta-se o SIFT, um método bem estabelecido na literatura e com diversas aplicações, incluindo reconhecimento de objetos e gestos, mapeamento e navegação robótica, modelagem 3D e monitoramento de vídeo. O descritor SIFT transforma uma imagem em uma coleção de vetores de características locais, sendo cada um deles invariante a translação, escala e rotação e parcialmente invariante a mudanças na iluminação ou projeção 3D ([LOWE, 1999](#)). Tais vetores estão localizados tanto no domínio espacial quanto no domínio das frequências, reduzindo a probabilidade de perturbações por oclusão, desorganização ou ruídos ([LOWE, 2004](#)).

A seguir são apresentados, segundo [Lowe \(2004\)](#), os quatro principais estágios de computação usados para gerar o conjunto de características de uma imagem.

1. **Detecção de extremos escala-espço.** Utiliza uma abordagem de filtragem em cascata, ilustrada na Figura 6a. A imagem na base da pilha à esquerda é a original, que passa pelo processo de convolução com filtros gaussianos. O uso do filtro gaussiano tem o objetivo de gerar imagens onde detalhes indesejados e ruídos são suavizados ou eliminados e características marcantes são realçadas. Então, imagens filtradas adjacentes são subtraídas para produzir as Diferenças de Gaussianas (pilha à direita na Figura 6a). O conjunto

composto pelas imagens filtradas e subtraídas constitui uma oitava. Após a oitava ser processada, a imagem filtrada com valor 2σ é reamostrada para metade de seu tamanho, dando início a uma nova oitava. Para detectar máximos e mínimos locais, cada *pixel* é comparado a seus oito vizinhos na imagem atual e nove vizinhos na escala acima e abaixo, como ilustrado na Figura 6b. O *pixel* é selecionado para a próxima etapa somente se possuir valor maior do que todos os seus vizinhos ou menor que todos eles. A invariância a mudanças de escala é alcançada por meio da busca em diferentes escalas.

Figura 6 – Ilustração da etapa “Detecção de extremos escala-espço” do descritor SIFT.

Fonte: Lowe (2004).

2. **Localização dos pontos-chave.** Os extremos detectados anteriormente e candidatos a pontos-chave são analisados visando eliminar os que possuem baixo contraste e, portanto, são sensíveis a ruídos e os que estão mal localizados, como em bordas.
3. **Atribuição de orientação.** Uma ou mais orientações são atribuídas a cada ponto-chave baseadas nas direções de gradientes locais da imagem. Todas as futuras operações são realizadas nos dados da imagem que foram transformados com relação a orientação, escala e localização atribuídos para cada característica, provendo assim invariância a essas transformações.
4. **Descritor de pontos-chave.** Os gradientes locais da imagem são medidos, na escala selecionada, na região ao redor de cada ponto-chave. Eles são transformados em uma

representação que permite níveis significativos de distorções locais de forma e mudanças na iluminação.

Um importante aspecto do descritor SIFT é que ele gera um grande número de vetores de características, de 128 dimensões, que cobrem a imagem em termos de escalas e localizações. Por exemplo, uma imagem de 500x500 *pixels* dará origem a cerca de 2000 vetores (embora esse número dependa tanto do conteúdo da imagem quanto da escolha de diversos parâmetros). A quantidade de características é particularmente importante para o reconhecimento de objetos, pois, para uma identificação confiável de pequenos elementos em fundos desordenados, pelo menos três características de cada objeto devem apresentar correspondência (LOWE, 2004).

Porém, a alta dimensionalidade dos vetores obtidos mostra-se uma desvantagem no passo de comparação de representações de diversas imagens (BAY; TUYTELAARS; VAN GOOL, 2006). Além disso, apesar do algoritmo SIFT ser considerado o mais eficaz, sem a ajuda de hardwares para acelerar a execução, computadores comuns dificilmente conseguem realizar o processamento em tempo real (HUANG *et al.*, 2011).

2.2.3 Medidas de similaridade

Obtidas as representações dos segmentos, são necessários mecanismos que permitam comparações consistentes. Uma possibilidade é utilizar medidas de similaridade, que expressam a relação entre representações vetoriais (MANNING; RAGHAVAN; SCHÜTZE, 2009). Nas Equações 2.3 e 2.4 apresentam-se dois exemplos: distância Euclidiana e cosseno, respectivamente.

A distância Euclidiana é a medida mais comumente utilizada e pode ser explicada como o comprimento da linha que conecta dois pontos. É adequada quando se trabalha com dados de baixa dimensionalidade e é importante considerar a magnitude dos vetores. Trata-se de uma medida intuitiva e simples de ser implementada.

$$D(A, B) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2} \quad (2.3)$$

O cosseno, por sua vez, calcula o ângulo cosseno entre dois vetores e é frequentemente utilizado com dados de alta dimensionalidade. Nesse caso, a magnitude não é considerada relevante por se tratar de uma medida de orientação. Dois vetores com a mesma direção possuem cosseno igual a 1, enquanto vetores opostos têm valor -1 (GROOTENDORST, 2021).

$$\cos(A, B) = \frac{\sum_{i=1}^n a_i b_i}{\sqrt{\sum_{i=1}^n a_i^2} \sqrt{\sum_{i=1}^n b_i^2}} \quad (2.4)$$

Outra opção são técnicas que avaliam a correspondência entre vetores de características. Uma dessas técnicas, que se destaca pelos resultados alcançados, é descrita a seguir.

One to One Symmetric matching (OOS)

Existem diversos algoritmos para encontrar correspondências entre vetores de características pertencentes a diferentes representações. Dependendo das restrições impostas, eles podem ser categorizados em muitos para muitos (M2M, de *many to many*), muitos para um (M2O, de *many to one*), um para muitos (O2M, de *one to many*) e um para um (O2O, de *one to one*). Os fatores que afetam a escolha da estratégia a ser utilizada incluem tolerância a ruídos, eficácia e eficiência da correspondência. No caso de vídeos, os quadros podem ter problemas de baixa resolução, perda de foco por movimentação e artefatos de compressão. Por isso, a tolerância a ruídos é um fator essencial na seleção do método de correspondência a ser utilizado (ZHAO *et al.*, 2007).

Nesse contexto, a técnica OOS (ZHAO *et al.*, 2007) destaca-se por ter como principal objetivo a otimização das correspondências e por apresentar desempenho superior a outros métodos encontrados na literatura. Isso porque a OOS é capaz de remover um grande número de correspondências causadas por ruídos e outras ambiguidades visuais presentes nas imagens (ZHOU *et al.*, 2009).

A técnica OOS é classificada como O2O, isto é, cada vetor de características em um conjunto pode ser mapeado a, no máximo, um vetor em outro conjunto. Além disso, assegura-se que vetores em um par sejam vizinhos mais próximos. O método também possui a propriedade de simetria, responsável por gerar resultados estáveis e únicos, uma vez que garante que a correspondência de um conjunto A para um conjunto B seja a mesma de B para A , a menos que os vetores de características possuam mais do que um vizinho mais próximo.

Outra característica da técnica OOS é permitir a correspondência parcial, ou seja, apenas um subconjunto de vetores de características é mapeado, a fim de se excluir pares com baixa similaridade. Para isso, adota-se um limiar, como indicado na Equação 2.5.

$$Sim(Q, P) \geq \alpha \quad (2.5)$$

O limiar α especifica a menor similaridade aceitável entre dois vetores de características para uma possível correspondência. A similaridade de dois vetores denotados por $Q = (q_1, q_2, \dots, q_n)$ e $P = (p_1, p_2, \dots, p_n)$, ou seja, $Sim(Q, P)$ (Equação 2.5), é dada pelo valor do ângulo cosseno entre eles.

Supondo K um conjunto de pares de vetores de características, que satisfazem a Equação 2.5 e foram identificados pela aplicação da técnica OOS em duas imagens, tem-se que a similaridade entre essas duas imagens pode ser determinada diretamente com base na cardinalidade de K ou considerando-se cada par de vetores, de acordo com a Equação 2.6 (ZHAO *et al.*, 2007).

$$KSim(K) = \frac{\sum_{(Q,P) \in K} Sim(Q, P)}{|K|} \quad (2.6)$$

onde:

$|K|$ denota a cardinalidade de K ;

$KSim$ calcula a similaridade média dos pares de vetores de características.

2.3 Sumarização de vídeo

Os métodos tradicionais de acesso a vídeos são ininterruptos e seriais: os usuários devem assistir ao fluxo de vídeo completo para conhecer seu conteúdo. Entretanto, esse processo é custoso em termos de tempo e recursos. A sumarização de vídeo possibilita que o conteúdo seja apresentado de maneira alternativa aos métodos tradicionais, oferecendo oportunidades para reduzir tempo e recursos exigidos do usuário (MONEY; AGIUS, 2008).

As duas principais aplicações da sumarização de vídeo são adaptação e navegação. A adaptação de vídeo busca transformar um ou mais vídeos de entrada por meio da manipulação de sinal, estrutura ou níveis semânticos, para satisfazer diversas restrições de recursos e preferências de usuário, ao mesmo tempo em que aperfeiçoa a utilidade do vídeo como um todo (CHANG; VETRO, 2005). A sumarização, portanto, pode ser vista como suporte ao processo de adaptação, por realizar uma transformação que retém apenas os elementos mais importantes (segundo algum critério) do conteúdo original. A navegação, por sua vez, visa aprimorar os meios utilizados pelos usuários para assistir a um fluxo de vídeo. A abordagem tradicional utilizada em reprodutores de mídia, como Windows Media Player e QuickTime, é baseada no acesso serial. Já a sumarização provê meios para o conteúdo de vídeo ser acessado de maneira visual e não linear, possibilitando que o usuário navegue pelo vídeo mais eficientemente (MONEY; AGIUS, 2008).

2.3.1 Processo de sumarização

O processo de sumarização pode ocorrer por meio da eliminação de dados temporais ou espaciais de um vídeo. No caso da sumarização temporal, o conteúdo é reduzido para uma versão de menor duração por meio da seleção de alguns segmentos, de acordo com critérios semânticos (NAPHIDE; HUANG, 2001). Em determinados casos, restrições de tempo são impostas, por exemplo, um sumário com os momentos mais violentos de um filme, mas com no máximo 2 minutos (MAGALHÃES; PEREIRA, 2004). Na Figura 7, apresenta-se um exemplo de sumarização temporal em que se simula a geração de dois sumários, sendo cada um deles formado por alguns segmentos do vídeo. É importante ressaltar que esses segmentos são trechos do vídeo, compostos pelo conteúdo audiovisual.

Já a sumarização espacial envolve avaliar cada quadro e remover alguns de seus elementos ou atribuir qualidade/resolução mais baixa a eles de acordo com a relevância semântica (MAGALHÃES; PEREIRA, 2004). Na Figura 8, ilustra-se a sumarização espacial, exemplificando um caso em que vários elementos foram removidos da cena, de maneira a torná-la menos

complexa.

Figura 7 – Exemplo de sumarização temporal.

Fonte: Adaptada de Magalhães e Pereira (2004).

Figura 8 – Exemplo de sumarização espacial.

Fonte: Magalhães e Pereira (2004).

2.3.2 Sumários de vídeo

Vale ressaltar que o foco deste trabalho é a sumarização temporal. Nesse contexto, os sumários de vídeo podem ser categorizados, segundo Money e Agius (2008), em: baseados em objetos, baseados em eventos, baseados em percepção e/ou baseados em características, dependendo do tipo de análise e dos critérios utilizados para se determinar o conteúdo a ser incluído. Além disso, os sumários podem ser diferenciados pela interatividade e personalização. Tais classificações estão ilustradas na Figura 9.

A seguir discutem-se as classificações baseadas no tipo de conteúdo analisado.

- **Sumários baseados em objetos:** concentram-se em objetos específicos que ocorrem em um vídeo, como uma cadeira, mesa, personagens principais, jogadores, entre outros. Sumários são incluídos nesta categoria se objetos são identificados durante a análise do vídeo e posteriormente usados como base para a formulação do sumário.

Figura 9 – Classificações de sumários de vídeo.

Fonte: Adaptada de Money e Agius (2008).

- **Sumários baseados em eventos:** identificam eventos durante o processo de análise e os utilizam para a criação dos sumários. Em vídeos de esporte, os eventos podem ser gols, cobranças de faltas, pênaltis, etc., já para filmes, uma briga ou cena de amor.
- **Sumários baseados em percepção:** derivam de conceitos de alto nível, representando os modos pelos quais o usuário compreende ou pode compreender o conteúdo do vídeo. Exemplos incluem o grau de entusiasmo que o usuário demonstra ao assistir o conteúdo, o nível de importância que associa ao conteúdo, a intensidade ou os tipos de emoção que pode observar no conteúdo ou ainda a atenção que determinados segmentos de vídeo demandam do usuário.
- **Sumários baseados em características:** são gerados de maneira objetiva a partir da análise de características de baixo nível, como mudanças na cor, textura e movimento nas imagens ou segmentos de silêncio ou fala no áudio. Vale destacar que, neste tipo de sumário, não se inferem objetos, eventos ou semântica baseada em percepções das características de baixo nível. Como resultado, os sumários são tipicamente apresentados como uma série de quadros-chave ou segmentos de vídeo sem descrições que esclareçam seu significado ou porque foram selecionados para representar o conteúdo sumarizado. Logo, pode-se dizer que tais sumários são os menos significativos semanticamente.

Como já mencionado, os sumários de vídeo também podem ser classificados em interativos ou estáticos e genéricos ou personalizados. **Sumários interativos** oferecem ao usuário diversos meios para interagir com o conteúdo, como navegar pelo sumário ou recuperá-lo de acordo com critérios semânticos expressos por meio de consultas. Essas consultas podem ser por palavras-chave ou listas de descritores textuais de objetos e eventos pelos quais o vídeo deve ser sumarizado. Outra possibilidade é utilizar imagens ou segmentos de vídeo de exemplo para se recuperar quadros-chave ou segmentos compatíveis. Já **sumários estáticos** apresentam uma série de quadros-chave ou segmentos estáticos ou ainda uma sequência de tempos indicando os segmentos de vídeo pertinentes.

Sumários personalizados, por sua vez, são desenvolvidos com base no entendimento individual que cada usuário possui do conteúdo. Tem se tornado cada vez mais reconhecido o fato de que os sumários devem ser personalizados para refletir a compreensão individual do conteúdo do vídeo, de maneira a serem intuitivos e considerarem necessidades particulares (LEW *et al.*, 2006; SILVA; YAMASAKI; AIZAWA, 2005). A personalização pode ser realizada de vários modos, por exemplo, filtrando-se o conteúdo do vídeo antes de ser apresentado ao usuário com o uso de perfis (TSENG; LIN; SMITH, 2004), sumarizando-se o conteúdo com base no comportamento e movimentos do usuário quando o vídeo é capturado (AIZAWA *et al.*, 2004) e utilizando-se padrões de acesso e navegação individuais (YU *et al.*, 2003). Em contraste, os **sumários genéricos** não são elaborados considerando-se a compreensão particular do conteúdo por cada usuário.

É importante destacar que, como a personalização deve incorporar o usuário durante o processo de análise, isto é, antes do sumário ser apresentado para utilização e consultas, os sumários que permitem que o usuário busque por conteúdos do vídeo não se qualificam automaticamente como personalizados. Por isso, apesar de diversos sumários incluírem uma função para se especificar a duração desejada, o resultado é classificado como genérico, a menos que os critérios para inclusão e exclusão de informações sejam derivados diretamente da compreensão do vídeo pelo usuário. Isso porque, muitas técnicas que permitem a definição de tempo como um parâmetro de entrada incluem e excluem conteúdo por meio de critérios genéricos e pré-definidos, sem considerar preferências individuais.

2.4 Avaliação de sumários de vídeo

Para se obter avanços em uma área de pesquisa, a eficácia de uma nova solução elaborada para resolver um problema específico precisa ser avaliada, preferencialmente em comparação a métodos já existentes. Para isso, pode ser necessária a consulta a usuários. Exemplos: 1) em um sistema de recuperação de informações - quais seriam os resultados corretos para a consulta “jaguar preto”? Carros ou animais? Cabe ao usuário decidir; 2) no contexto de sumarização de vídeo - qual o melhor entre os resultados produzidos por técnicas diferentes? A resposta depende da percepção do usuário.

Entretanto, desenvolver estudos com indivíduos reais é uma tarefa custosa, difícil de controlar e replicar. Por essa razão, em diversas situações, podem ser utilizadas as chamadas coleções de testes. Em geral, tais coleções são criadas com o envolvimento dos usuários, mas uma vez disponíveis, um novo método pode ser avaliado em um experimento controlado (BLANKEN *et al.*, 2007). Assim, nesta seção, apresentam-se medidas objetivas (Seção 2.4.1) e protocolos de avaliação subjetiva (Seção 2.4.2) utilizados na área de sumarização multivídeo.

2.4.1 Avaliação objetiva

Uma metodologia de avaliação objetiva envolve três itens (BLANKEN *et al.*, 2007):

1. uma coleção de testes, consistindo em dados multimídia, uma tarefa e informações necessárias para realizá-la (por exemplo, busca por imagens usando imagens representativas como consultas) e um *ground truth*, também chamado de julgamentos de relevância, isto é, as respostas corretas;
2. uma ou mais medidas de avaliação;
3. uma metodologia estatística para determinar se as diferenças observadas no desempenho dos métodos investigados são estatisticamente significativas.

Com o objetivo de compartilhar recursos, as coleções de testes (item 1) são frequentemente desenvolvidas em um esforço colaborativo de diversos grupos de pesquisa, como em *workshops* de conferências. Usuários são envolvidos no processo de criação de *ground truths*. Por exemplo, para a avaliação de um algoritmo de detecção de tomadas, são feitas anotações manuais indicando se uma sequência de quadros contém ou não uma transição (BLANKEN *et al.*, 2007).

Algumas métricas (item 2) que se originaram na área de recuperação de informação, podem ser adaptadas a outros contextos, como sumarização. Nesse caso, as consultas podem ser mapeadas aos sumários de vídeo e os documentos recuperados aos segmentos incluídos. Para entender as equações que serão apresentadas, é necessário primeiramente definir alguns termos. Assim, tem-se que:

- verdadeiros positivos (vp) são documentos que foram recuperados e são relevantes;
- verdadeiros negativos (vn) são documentos que não foram recuperados e não são relevantes;
- falsos positivos (fp) são documentos que foram recuperados, mas não são relevantes;
- falsos negativos (fn) são documentos que não foram recuperados, mas são relevantes.

Na Tabela 1, mostram-se as definições apresentadas anteriormente de forma resumida.

Tabela 1 – Definições de verdadeiros positivos e negativos, falsos positivos e negativos.

	Relevantes	Não relevantes
Recuperados	Verdadeiros positivos (vp)	Falsos positivos (fp)
Não recuperados	Falsos negativos (fn)	Verdadeiros negativos (vn)

Fonte: Manning, Raghavan e Schütze (2009).

Precisão e abrangência (do inglês *recall*, também traduzida como cobertura ou revocação) são as duas medidas mais frequentemente utilizadas para aferir a eficácia de sistemas de recuperação de informações. A seguir, tais medidas são apresentadas, segundo [Manning, Raghavan e Schütze \(2009\)](#).

A precisão (P) é definida como a fração de documentos recuperados que são relevantes.

$$P = \frac{n.^{\circ} \text{ itens relevantes recuperados}}{n.^{\circ} \text{ itens recuperados}} \quad (2.7)$$

$$P = \frac{vp}{vp + fp} \quad (2.8)$$

A abrangência (A), por sua vez, é definida como a fração de documentos relevantes que foram recuperados.

$$A = \frac{n.^{\circ} \text{ itens relevantes recuperados}}{n.^{\circ} \text{ itens relevantes}} \quad (2.9)$$

$$A = \frac{vp}{vp + fn} \quad (2.10)$$

Outra forma de avaliação comumente utilizada é a medida F, que compara precisão e abrangência usando a média harmônica ponderada. De acordo com [Manning, Raghavan e Schütze \(2009\)](#), é definida como:

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{A}} = \frac{(\beta^2 + 1)PA}{\beta^2 P + A}, \text{ onde } \beta^2 = \frac{1 - \alpha}{\alpha} \quad (2.11)$$

sendo $\alpha \in [0, 1]$ e $\beta \in [0, \infty]$.

A medida F padrão atribui o mesmo peso para precisão e abrangência, o que significa $\alpha = \frac{1}{2}$ ou $\beta = 1$. Essa medida é comumente descrita como F_1 , que é uma abreviação para $F_{\beta=1}$. Neste caso, a Equação 2.11 é simplificada para:

$$F_1 = \frac{2PA}{P + A} \quad (2.12)$$

O cálculo da medida F, a qual utiliza média harmônica, mostra-se mais interessante do que o cálculo de uma simples média aritmética entre precisão e abrangência, por exemplo. Isso porque é sempre possível alcançar 100% de abrangência (basta que todos os documentos sejam retornados). Considerando-se essa situação, ao calcular a média aritmética sempre seria possível atingir, no mínimo, a taxa de 50%, o que indica que o uso de tal medida é inadequado. Já a média harmônica é sempre menor ou igual à média aritmética e à média geométrica. Quando dois números diferem muito, a média harmônica é mais próxima ao menor dos valores do que à média aritmética de ambos.

Por fim, testes de significância estatística (item 3) são importantes para mostrar que as diferenças entre dois métodos não se devem a variações aleatórias de desempenho. Exemplos incluem: teste-T, Wilcoxon e teste do sinal. Para utilizar o teste-T os erros devem possuir distribuição normal e para o Wilcoxon precisam ser contínuos. Caso contrário, o teste do sinal pode ser aplicado como uma alternativa (BLANKEN *et al.*, 2007).

2.4.2 Avaliação subjetiva

Nesta seção, discutem-se duas abordagens de avaliação subjetiva de sumários multivídeo, utilizadas em trabalhos relacionados: 1) matriz de preferências e 2) baseada em notas. A primeira delas se destaca por possibilitar uma visão global dos resultados: em comparações par a par, os usuários expressam suas preferências em um cenário que avalia todos contra todos. Na segunda abordagem, os participantes do experimento indicam quão satisfeitos estão com cada sumário apresentado por meio de uma nota, fornecendo uma medida para as análises.

Matriz de preferências

Nesta abordagem, os sumários multivídeo obtidos por métodos diferentes são apresentados em pares e os usuários fazem escolhas entre as opções.

Figura 10 – Exemplo de resultado da abordagem de avaliação subjetiva “Matriz de preferências”.

	OCFE	DSVS	MKFE	SRSB	PM	Score	Rank
OCFE	--	8	6	17	7	38	5
DSVS	22	--	8	14	15	59	4
MKFE	24	22	--	10	14	70	2
SRSB	13	16	20	--	11	60	3
PM	23	15	16	19	--	73	1

The Merlion

	OCFE	DSVS	MKFE	SRSB	PM	Score	Rank
OCFE	--	6	3	11	9	29	5
DSVS	24	--	13	13	17	67	3
MKFE	27	17	--	12	11	67	2
SRSB	19	17	18	--	9	63	4
PM	21	13	19	21	--	84	1

The Marina Bay

	OCFE	DSVS	MKFE	SRSB	PM	Score	Rank
OCFE	--	2	4	12	6	24	5
DSVS	28	--	17	15	12	72	2
MKFE	26	13	--	8	8	55	4
SRSB	18	15	22	--	11	66	3
PM	24	18	22	19	--	83	1

The Esplanade

	OCFE	DSVS	MKFE	SRSB	PM	Score	Rank
OCFE	--	5	8	14	7	50	5
DSVS	25	--	15	11	16	67	2
MKFE	22	15	--	14	10	61	4
SRSB	16	19	16	--	14	65	3
PM	23	14	20	16	--	73	1

The Singapore Flyer

Fonte: Zhang *et al.* (2015a).

Na Figura 10, é possível observar um exemplo de apresentação dos resultados da avaliação (ZHANG *et al.*, 2015a). Nesse caso, o método proposto foi comparado a outros 4 da literatura, por 30 usuários, para 4 conjuntos de vídeos. Cada elemento da matriz traz a informação de quantas vezes um método foi escolhido em comparação a outro, considerando linha *versus* coluna. Por exemplo, o valor contido na linha 1, coluna 2 da matriz *The Marina Bay* indica que 6 participantes preferiram o método OCFE ao invés do DSVS. A soma dos elementos de cada linha fornece a pontuação total obtida (coluna *Score*), possibilitando a ordenação dos métodos de acordo com a opinião dos usuários

(coluna *Rank*). Cada matriz corresponde aos resultados da avaliação para um conjunto de vídeos sumarizado.

Baseada em notas

Nesta abordagem, de acordo com os trabalhos relacionados investigados, os usuários atribuem notas, em geral, de 1 (ruim) a 10 (bom) aos resultados gerados por diferentes técnicas de sumarização multivídeo. São passadas orientações quanto ao que deve ser analisado nos sumários, as quais dependem do contexto. Por exemplo, [Zhang e Zimmermann \(2016\)](#) solicitaram que os participantes considerassem a relevância do conteúdo e a qualidade visual ao definir as notas. [Ji et al. \(2018\)](#), por sua vez, propuseram que a avaliação levasse em conta as seguintes propriedades: importância, cobertura, adaptabilidade e entendimento.

Na Figura 11, pode-se observar um exemplo de apresentação dos resultados ([ZHANG; ZHANG; ZIMMERMANN, 2015](#)). Nesse caso, a avaliação envolveu 19 usuários, que atribuíram notas de 1 a 10 aos sumários produzidos por 3 técnicas, para 6 conjuntos de vídeos. Na Figura 11a, indica-se a média das notas e desvio padrão por conjunto e método. Na Figura 11b, mostram-se as preferências individuais, ou seja, quantas vezes (em porcentagem) os participantes preferiram os resultados de cada técnica de sumarização multivídeo. Outra análise usualmente realizada em trabalhos relacionados que adotam essa abordagem de avaliação subjetiva é o cálculo da média das notas, por usuário, para os sumários de cada técnica. Na Figura 12, traz-se um exemplo extraído de ([JI et al., 2018](#)), que apresenta a nota média de cada um dos 5 usuários para os resultados de 6 métodos.

Figura 11 – Exemplo de resultado da abordagem de avaliação subjetiva baseada em notas.

	Média $\pm$ Desvio padrão		
	AMVS	SC	MMR
conj.(1)	7.842 $\pm$ 0.933	4.579 $\pm$ 1.270	3.632 $\pm$ 1.265
conj.(2)	7.526 $\pm$ 0.939	5.158 $\pm$ 0.812	4.421 $\pm$ 1.600
conj.(3)	7.632 $\pm$ 0.809	4.316 $\pm$ 1.656	4.632 $\pm$ 1.529
conj.(4)	8.421 $\pm$ 0.494	6.158 $\pm$ 0.744	4.000 $\pm$ 1.919
conj.(5)	7.842 $\pm$ 0.987	5.316 $\pm$ 1.524	5.053 $\pm$ 1.276
conj.(6)	7.684 $\pm$ 1.300	6.316 $\pm$ 1.259	5.737 $\pm$ 1.996

- (a) Média e desvio padrão das notas atribuídas pelos usuários aos resultados de 3 técnicas, para 6 conjuntos de vídeos.
- (b) Preferências individuais: quantas vezes (em porcentagem) os usuários preferiram os resultados de cada técnica.

Fonte: [Zhang, Zhang e Zimmermann \(2015\)](#).

Figura 12 – Exemplo de resultado da abordagem de avaliação subjetiva baseada em notas. Apresenta-se a média das notas atribuídas, por usuário, para os resultados de cada técnica.

Fonte: Ji *et al.* (2018).

2.5 Considerações finais

Neste capítulo, foram apresentados conceitos relacionados ao contexto do trabalho, provendo uma visão ampla da área de pesquisa, por meio de tópicos como as propriedades de um vídeo digital e também aprofundada, por exemplo, com a definição de extratores de características apropriados ao domínio e descrição de abordagens de avaliação de técnicas de sumarização multivídeo. No próximo capítulo, apresentam-se uma análise da literatura, evidenciando seu estágio atual, bem como lacunas de pesquisa.

TRABALHOS RELACIONADOS

Um passo essencial no processo de sumarização é a definição de critérios a serem aplicados para avaliar a relevância dos segmentos de vídeo e decidir quais devem ser incluídos no sumário.

Vale destacar que a produção de sumários a partir de um ou múltiplos vídeos compartilha o mesmo objetivo: selecionar os segmentos mais importantes dos vídeos de entrada e gerar uma versão representativa de menor duração. Entretanto, a sumarização multivídeo é um problema mais complexo e adiciona novos desafios à tradicional utilização de uma única fonte ([JI *et al.*, 2018](#); [PANDA](#); [MITHUN](#); [ROY-CHOWDHURY, 2017](#)). Essa área de pesquisa precisa lidar com conteúdo redundante no conjunto de entrada e, ao mesmo tempo, explorar informações complementares para enriquecer os resultados. Além disso, como apontado por [Panda, Mithun e Roy-Chowdhury \(2017\)](#), outros desafios no processo de sumarização incluem fatores ambientais como diferenças de iluminação, ângulo e problemas de sincronização entre múltiplos vídeos relacionados. [Ji *et al.* \(2018\)](#) acrescentam que é difícil definir uma ordem de apresentação lógica, uma vez que os quadros ou tomadas selecionados provêm de diferentes fontes.

Portanto, soluções projetadas para obter um sumário a partir de um vídeo não podem ser diretamente aplicadas para sumarizar um conjunto. Assim, neste capítulo, o foco é a análise de técnicas de sumarização multivídeo, em especial, os critérios por elas utilizados para seleção de conteúdo.

3.1 Protocolo para o mapeamento sistemático

Desenvolveu-se um mapeamento sistemático da literatura ([KITCHENHAM, 2004](#)), atualizado em 07/01/2021. A pesquisa foi realizada na ferramenta Google Acadêmico com a seguinte *string* de busca: “*multi-video summarization*” OR “*multi-video synopsis*” OR “*multi-document video summarization*” OR “*multi-document video synopsis*”.

Optou-se por analisar documentos provenientes das seguintes bases de dados indexadas: *IEEE Xplore*, *ACM Digital Library*, *SpringerLink* e *ScienceDirect*, por serem fontes confiáveis de informação e amplamente utilizadas na comunidade científica. Definiu-se como critério de inclusão: trabalhos relacionados a sumarização multivídeo e como critérios de exclusão, trabalhos: 1) que não estivessem escritos em português ou inglês; 2) sem PDF disponível; 3) não pertencentes às bases de dados previamente selecionadas; 4) não relacionados à sumarização multivídeo.

No total, foram retornados 197 artigos. Eles foram submetidos à etapa de seleção preliminar, que consistiu na leitura dos títulos e resumos para avaliar se satisfaziam os critérios de inclusão/exclusão estabelecidos. Então realizou-se a seleção final, ou seja, a leitura completa para definição de quais deveriam permanecer aceitos. Na Tabela 2 indica-se o número de trabalhos excluídos por cada critério sendo que, ao final do processo, 29 foram marcados para análise.

Tabela 2 – Número de trabalhos excluídos por cada critério durante as etapas de seleção preliminar e seleção final do mapeamento sistemático.

Critério de exclusão	Número de trabalhos
Idioma diferente de português ou inglês	10
Sem PDF disponível	7
Não pertencentes às bases de dados selecionadas	46
Não relacionados a sumarização multivídeo	105

Fonte: Dados da pesquisa.

3.2 Resultado do mapeamento sistemático

Como um primeiro resultado da análise dos trabalhos selecionados pelo processo descrito na Seção 3.1, apresenta-se um resumo da literatura descrevendo-se, em geral, os critérios para seleção de conteúdo identificados.

Redundância. A importância de um segmento é avaliada pela repetição de conteúdo, isto é, ele é selecionado se representa informações que aparecem mais de uma vez no conjunto de vídeos.

Cobertura. Um segmento é incluído no sumário se é capaz de tornar o resultado mais completo, cobrindo o máximo possível de informações relevantes.

Qualidade/estética. A relevância de um segmento é mensurada considerando aspectos estéticos, como condições de iluminação, foco, cor e textura.

Diversidade. Um segmento é selecionado para compor o sumário caso possua conteúdo diferente dos que já foram incluídos, aumentando a diversidade de informações apresentadas aos usuários.

Consistência visual/coerência. As transições entre os segmentos selecionados devem ser suaves, de modo que não prejudiquem a experiência de visualização do usuário.

Objetos em movimento. Um segmento é considerado relevante para ser incluído no sumário se possui objetos em movimento.

Cronologia. Segmentos de diferentes fontes são inseridos no sumário considerando a ordem cronológica, isto é, provendo um enredo coerente e um resultado mais compreensível.

Na Tabela 3, quantifica-se o uso de cada critério nos trabalhos relacionados. Uma análise superficial indica que a maioria das técnicas de sumarização multivídeo é baseada em redundância e diversidade. Em outras palavras, esses trabalhos identificam informações comuns nos diferentes vídeos, com a preocupação de não incluir conteúdo repetido.

Tabela 3 – Critérios para seleção de conteúdo identificados em trabalhos relacionados a sumarização multivídeo.

Critério	Número de trabalhos (%)
Redundância	58,6
Cobertura	13,8
Qualidade/estética	20,7
Diversidade	58,6
Consistência visual/coerência	17,2
Objetos em movimento	3,5
Cronologia	6,9

Fonte: Dados da pesquisa.

3.3 Análise dos trabalhos relacionados

3.3.1 Critérios para seleção de conteúdo

Mais do que uma visão geral, apresentada na Seção 3.2, é importante desenvolver uma análise mais aprofundada para entender qual a definição dos critérios em contextos de uso específicos e como são aplicados para a produção de sumários.

Na Tabela 4 apresentam-se os critérios para seleção de conteúdo empregados em cada trabalho investigado. Nas técnicas de sumarização multivídeo reportadas em (HONG *et al.*, 2009; WANG; Merialdo, 2009; DALE *et al.*, 2012; WANG *et al.*, 2012; IDE *et al.*, 2017; ZHANG *et al.*, 2020), a geração de um sumário relevante é baseada em redundância, isto é, identificação de sobreposição de conteúdo em diferentes vídeos. Para aplicar esse critério, Hong *et al.* (2009) extraem quadros-chave das tomadas de vídeo e cada um deles é representado por características visuais locais (SIFT). Se pares de quadros-chave de diferentes vídeos possuem valor de similaridade acima de um limiar pré-definido, essas tomadas são rotuladas como duplicadas. Wang e

Tabela 4 – Comparação entre os trabalhos relacionados quanto aos critérios para seleção de conteúdo.

	Redundância	Cobertura	Qualidade	Diversidade	Coerência	Movimento	Cronologia
Hong <i>et al.</i> (2009)	✓	✗	✗	✗	✗	✗	✗
Wang e Merialdo (2009)	✓	✗	✗	✗	✗	✗	✗
Dale <i>et al.</i> (2012)	✓	✗	✗	✗	✗	✗	✗
Wang <i>et al.</i> (2012)	✓	✗	✗	✗	✗	✗	✗
Ide <i>et al.</i> (2017)	✓	✗	✗	✗	✗	✗	✗
Zhang <i>et al.</i> (2020)	✓	✗	✗	✗	✗	✗	✗
Li e Merialdo (2010a)	✓	✗	✗	✓	✗	✗	✗
Li e Merialdo (2010b)	✓	✗	✗	✓	✗	✗	✗
Li e Merialdo (2011)	✓	✗	✗	✓	✗	✗	✗
Li e Merialdo (2012)	✓	✗	✗	✓	✗	✗	✗
Li e Merialdo (2016)	✓	✗	✗	✓	✗	✗	✗
Li <i>et al.</i> (2011)	✓	✗	✗	✓	✗	✗	✗
Ou <i>et al.</i> (2014)	✓	✗	✗	✓	✗	✗	✗
Kanehira <i>et al.</i> (2018)	✓	✗	✗	✓	✗	✗	✗
Wu, Zhong e Liu (2020)	✓	✗	✗	✓	✗	✗	✗
Ji <i>et al.</i> (2018)	✓	✗	✗	✓	✗	✗	✓
Ji <i>et al.</i> (2019)	✓	✗	✗	✓	✗	✗	✓
Ou <i>et al.</i> (2015)	✗	✗	✗	✓	✗	✗	✗
Panda, Mithun e Roy-Chowdhury (2017)	✗	✗	✗	✓	✗	✗	✗
Panda e Roy-Chowdhury (2017)	✗	✗	✗	✓	✗	✗	✗
Ji <i>et al.</i> (2020)	✗	✗	✗	✓	✗	✗	✗
Zhang, Ma e Zimmermann (2013)	✗	✓	✓	✓	✗	✗	✗
Zhang <i>et al.</i> (2012)	✗	✓	✗	✓	✓	✗	✗
Zhang, Zhang e Zimmermann (2015)	✗	✓	✓	✗	✓	✗	✗
Zhang e Zimmermann (2016)	✗	✓	✓	✗	✓	✗	✗
Zhang <i>et al.</i> (2015b)	✗	✗	✓	✗	✗	✗	✗
Zhang <i>et al.</i> (2015a)	✗	✗	✓	✗	✓	✗	✗
Nie <i>et al.</i> (2016)	✗	✗	✓	✗	✓	✗	✗
Zhu, Liao e Li (2016)	✗	✗	✗	✗	✗	✓	✗

Fonte: Dados da pesquisa.

Merrialdo (2009) também utilizam informações visuais para medir a similaridade entre quadros, combinando características globais de cor e textura (histogramas *Hue*, *Saturation* and *Value* (HSV) e *Haar wavelet*, respectivamente) com locais (SIFT).

Dale *et al.* (2012) compartilham o mesmo objetivo: encontrar quadros em comum no conjunto de entrada. Para isso, estimam um valor de representatividade, computando a similaridade entre um quadro de um vídeo específico e os quadros dos demais vídeos que compõem a coleção. Além disso, o método de Wang *et al.* (2012) é baseado na identificação de tomadas-chave, isto é, tomadas que contêm as informações principais de um evento. Para encontrá-las, os autores agrupam quadros-chave extraídos das tomadas e representados por características visuais e, então, detectam repetições em cada grupo.

Ide *et al.* (2017) propõem um método para sumarizar múltiplos vídeos de notícias, o qual também considera a redundância entre eles. A abordagem segue o raciocínio de que a seleção do

conteúdo de áudio mais importante (processado como texto) deve preceder a do conteúdo visual. Como entrada, eles esperam uma sequência temporalmente ordenada de trechos de notícias. Primeiramente, atribuem uma pontuação a cada sentença considerando que as relevantes contêm mais conceitos visuais presentes em um vocabulário pré-definido. Desse modo, a sentença com maior valor é selecionada como sentença-chave, representando cada trecho de notícia. Então, o vídeo é segmentado em tomadas usando histogramas de cor HSV e o primeiro quadro é extraído. Conceitos visuais são detectados nos quadros-chave das tomadas. Por fim, para cada trecho de notícia, a tomada cujo conteúdo visual é consistente com a sentença-chave correspondente é escolhida.

Zhang *et al.* (2020), por sua vez, introduzem um método no domínio de vigilância. Primeiramente, sincronizam os vídeos de entrada e identificam o mesmo objeto em diferentes fontes, agrupando-os. Então, os grupos são deslocados no tempo. Dessa maneira, são produzidos múltiplos sumários correspondentes a cada entrada. Para obter o resultado, propõem um algoritmo que encontra o melhor caminho de visualização entre os sumários individuais.

De modo similar, Li e Merialdo em suas diferentes versões de uma técnica de sumarização multivídeo acreditam que redundância e diversidade devem ser usadas em conjunto para atingir resultados significativos. Para eles, adotar esses critérios significa procurar por segmentos de vídeo com conteúdo similar a todo o conjunto, mas diferente dos segmentos já incluídos no sumário. Em (LI; Merialdo, 2010b; LI; Merialdo, 2016) - Video-MMR, os autores usam características visuais (SIFT e *Bag of Words*) para medir a similaridade entre quadros. Em (LI; Merialdo, 2010a; LI; Merialdo, 2016) - AV-MMR, adicionam informações de áudio e calculam a similaridade entre segmentos de vídeo combinando *Bag of Words* e *Mel-Frequency Cepstral Coefficients* (MFCC). Em uma terceira versão: (LI; Merialdo, 2012; LI; Merialdo, 2016) - BAV-MMR, os critérios são aplicados usando características de alto nível (detecção de faces e gênero de áudio) e informações temporais. Além disso, (LI; Merialdo, 2011) - OB-MMR estende BAV-MMR calculando os parâmetros do algoritmo automaticamente.

Adotando os mesmos critérios, isto é, redundância e diversidade, há ainda o trabalho de Li *et al.* (2011) que usa características visuais em conjunto com textuais. Os autores empregam histogramas de palavras visuais e realizam a transcrição de áudio com *Laboratoire d'Informatique d'Avignon, France - Automatic Speech Recognition* (LIA ASR). De modo similar, Ou *et al.* (2014) propõem uma adaptação da técnica Video-MMR (LI; Merialdo, 2010b) para processamento *on-line*. Nesse trabalho, os quadros são representados por histogramas de cor no espaço HSV.

Kanehira *et al.* (2018) argumentam que dados múltiplos grupos de vídeos, o sumário resultante deve ser diverso, representativo dos vídeos em um mesmo grupo e discriminativo contra vídeos em grupos diferentes. Portanto, o método é baseado em redundância e diversidade. Para atingir o objetivo, os autores propõem minimizar a variância intra-grupo e maximizar a variância entre grupos, inspirados por conceitos de análise discriminante linear. Nesse trabalho, aplica-se aprendizado de características; mais especificamente, exploram redes *Convolutional*

3D (C3D) como extratores de características.

Wu, Zhong e Liu (2020) também consideram redundância e diversidade. Os autores tratam as tomadas de vídeo como nós de um grafo e seus relacionamentos par a par como pesos de arestas. Utilizam uma rede convolucional para realizar a classificação dos nós e, então, para seleção de conteúdo seguem os seguintes passos: 1) criam um *ranking* com os nós rotulados com valor 1 em ordem decrescente de número de arestas conectadas a outros nós; 2) selecionam tomadas (nós) para o sumário, até atingir o tamanho desejado, com base na ordenação realizada anteriormente. Quando uma tomada é incluída no sumário, as similares são removidas. Utilizam conteúdo visual das tomadas: características C3D de 4096 dimensões e cosseno como medida de similaridade.

Ji *et al.* (2018) propõem uma técnica de sumarização multivídeo baseada principalmente em redundância, diversidade e também cronologia. Para esse propósito, introduzem um método de agrupamento *Hypergraph-based Dominant Set* (HDS) para descobrir padrões visuais, o qual é capaz de modelar relações de alta ordem entre múltiplos vídeos. Os centróides dos conjuntos dominantes são selecionados como quadros-chave candidatos. Então, utilizam imagens da web obtidas pela mesma consulta para identificar quadros-chave com similaridade: máxima a tais imagens e mínima aos quadros já selecionados. Para a apresentação do sumário, os autores desenvolveram uma abordagem baseada em grafo para gerar um enredo coerente, tornando o resultado mais compreensível. Os vídeos usados nos experimentos foram pré-processados com detecção de tomadas e o quadro mediano de cada uma delas foi extraído. Foram empregadas características visuais compostas por vetores *Convolutional Neural Network* (CNN) VGGNet-19 com 4096 dimensões e histogramas de cor HSV com 256 dimensões.

Baseando-se na mesma combinação de critérios (redundância, diversidade e cronologia), tem-se o trabalho de Ji *et al.* (2019). Primeiramente, um grafo é construído mostrando as relações e importância de quadros-chave candidatos. Isso é feito considerando-se características visuais e imagens obtidas da web. Então, os candidatos são agrupados aplicando-se um algoritmo de aprendizado não supervisionado sobre o grafo. Cria-se um *ranking* com tais grupos, em ordem decrescente de relevância, visando priorizar o conteúdo que será levado ao sumário. São selecionados dois quadros pertencentes aos grupos com melhor classificação e, para garantir diversidade, os pares são comparados entre si. Caso sejam considerados similares, um deles é excluído do resultado. Os autores também propõem um método para apresentar o sumário de maneira mais legível e lógica, o qual considera relações temporais entre os quadros-chave e proximidade quanto ao tópico abordado.

Concentrando-se em diversidade, Ou *et al.* (2015) filtram quadros redundantes entre vídeos diferentes. Eles computam um histograma de cor normalizado sobre o primeiro plano de uma imagem e um valor de importância considerando o tamanho do primeiro plano. Em um segundo passo, essas características extraídas dos quadros são comparadas para decidir quais serão removidos.

Panda, Mithun e Roy-Chowdhury (2017) e Panda e Roy-Chowdhury (2017) desenvolvem uma abordagem de sumarização multivídeo baseada no que denominam complementaridade. Esse critério pode ser mapeado ao de diversidade, já que o principal objetivo é evitar a seleção de segmentos similares provenientes de diferentes vídeos. Primeiramente, eles dividem cada vídeo em múltiplos segmentos não uniformes, medindo a quantidade de mudanças entre dois quadros consecutivos nos espaços de cor RGB e HSV. Então, os segmentos são representados usando características de aprendizado profundo. Mais precisamente, os autores extraem características C3D, usando conjuntos de entrada com 16 quadros, aplicando filtros convolucionais 3D e obtendo as respostas na camada FC6. Após um esquema de *pooling* médio temporal, o vetor de características final (4096 dimensões) é gerado. O próximo passo visa a produção do sumário em si. Para isso, eles propõem uma extensão de abordagens de otimização esparsa para selecionar segmentos que são interessantes e representativos. Para explorar informações complementares, os autores introduzem uma função regularizadora que penaliza a condição de que dois segmentos correlacionados de dois vídeos distintos sejam apresentados no sumário ao mesmo tempo.

Ji *et al.* (2020) também propõem um método que considera diversidade. Aplicam *sparse auto-encoder*, cujo objetivo é aprender automaticamente representações eficazes de dados, capturar seu conteúdo vital e reconstruir as informações originais. Nesse trabalho é empregado para identificação de quadros-chave, combinando informações visuais (características profundas extraídas com VGGNet-19 e histogramas HSV concatenados em um vetor de 4352 dimensões) e textuais (abordagem TF-IDF). Em um passo posterior, a similaridade (cosseno) entre quadros é calculada, visando remover repetições e garantir diversidade.

Por outro lado, há um conjunto de trabalhos (ZHANG; MA; ZIMMERMANN, 2013; ZHANG *et al.*, 2012; ZHANG; ZHANG; ZIMMERMANN, 2015; ZHANG; ZIMMERMANN, 2016) baseados em combinações dos seguintes critérios: cobertura, qualidade, diversidade e consistência visual.

Zhang, Ma e Zimmermann (2013) combinam cobertura, qualidade e diversidade. O critério cobertura visa selecionar segmentos de vídeo que mostram um objeto de diferentes pontos de vista porque, segundo os autores, cada um deles o apresenta com uma semântica específica. Para isso, calculam perspectiva angular e distância entre câmera e objeto. Em adição, a qualidade captura quão bem uma informação é armazenada. Tal critério é implementado usando um fator que considera com melhor qualidade objetos aparecendo em visão central ou distância mais próxima. Por fim, a diversidade evita a inclusão de conteúdo repetido no sumário. Isso é feito avaliando-se a similaridade entre segmentos usando valores de distância e perspectiva angular.

Zhang *et al.* (2012) também baseiam seu trabalho em cobertura e diversidade, considerando para esses critérios a mesma definição apresentada em (ZHANG; MA; ZIMMERMANN, 2013). Entretanto, ao invés de adotar qualidade, eles elegeram consistência visual para complementar o processo de geração de sumários, o que significa que as transições de segmentos devem

ser o mais suave possível, provendo aos usuários uma experiência de visualização agradável. Para alcançar os objetivos, os autores definem fatores para cada critério baseados em valores de ângulo e movimento.

Para [Zhang, Zhang e Zimmermann \(2015\)](#) é importante considerar: 1) cobertura: reflete quanta informação de destaque/importante dos vídeos originais permanece no sumário; 2) consistência visual: mesma definição de ([ZHANG et al., 2012](#)); 3) qualidade estética: significa que um bom sumário deve incluir os melhores segmentos, já que vídeos com baixa qualidade prejudicam a experiência e a satisfação dos usuários.

Nesse trabalho, os autores dividem cobertura em *inter-coverage*, isto é, verificam se um vídeo contém uma cena essencial considerando o conjunto de entrada completo e *intra-coverage*, que mede quão bem ele representa essa cena específica. Para consistência visual, avaliam a similaridade entre pares de quadros, calculando um valor de contribuição para manter o sumário consistente. Por fim, para qualidade estética, assumem que a percepção de um quadro é influenciada por seus vizinhos. Portanto, um fator é introduzido para capturar o nível de estímulo: uma grande surpresa atrai atenção, então a qualidade é percebida mais intensamente e nível de ruído: a audiência obtém menos informação de um vídeo se o fundo dos objetos principais se movimenta muito. Para computar qualidade para cada quadro, os autores usam características visuais: histogramas de cor, momentos de cor, *Histogram of Oriented Gradients* (HOG), *Local Binary Pattern* (LBP), *Difference of Gaussian* (DoG) e mapa de saliência. O nível de estímulo é obtido por fluxo óptico e o nível de ruído por movimento de câmera. Vale destacar que [Zhang e Zimmermann \(2016\)](#) aplicam os mesmos critérios discutidos para ([ZHANG; ZHANG; ZIMMERMANN, 2015](#)).

Seguindo a mesma linha, [Zhang et al. \(2015b\)](#) adotam estética. Na técnica, um quadro é descrito pelas características: composição global de matiz e composição global de cena. Os autores propõem um modelo probabilístico que usa tais características extraídas das imagens para identificar segmentos que melhor satisfazem o critério. O trabalho de [Zhang et al. \(2015a\)](#) combina estética, coerência e estabilidade (o que pode ser classificado como um aspecto de qualidade). O critério estética é avaliado usando características de cor e textura (momentos de cor e HOG, respectivamente). Nessa técnica, coerência está sendo usada como sinônimo de consistência visual já discutida em trabalhos anteriores, em outras palavras, a fluência de um sumário de vídeo. A medida de coerência é definida como um núcleo gaussiano. Por fim, estabilidade reflete o número de quadros desfocados. É calculada acumulando-se as diferenças entre quadros adjacentes, representados por características de cor e textura.

[Nie et al. \(2016\)](#) empregam os mesmos critérios mostrados em ([ZHANG et al., 2015a](#)). Porém, eles possuem uma estratégia diferente para medir estabilidade: é obtido um vetor de movimento de câmera de um quadro para outro e grandes oscilações em tais vetores indicam que um segmento de vídeo é instável.

Com uma abordagem diferente, [Zhu, Liao e Li \(2016\)](#) propõem como critério para seleção

de conteúdo a identificação de sequências de quadros com o mesmo objeto em movimento. Por isso, eles executam duas tarefas principais: 1) segmentação de objetos em movimento: característica *Scale-Invariant Local Ternary Pattern* (SILTP) com subtração de plano de fundo; 2) rastreamento de objetos: algoritmo que conecta o mesmo objeto em quadros diferentes. É importante destacar que esse critério está fortemente relacionado ao domínio da técnica: vídeos de vigilância.

3.3.2 Bases de vídeo

Além dos critérios para seleção de conteúdo, discutidos na Seção 3.3.1, realizou-se um levantamento das bases de vídeo utilizadas para avaliação das técnicas de sumarização multivídeo. Na Tabela 5, apresenta-se uma síntese dos resultados. Destaca-se que não há uma padronização quanto às características de cada conjunto de vídeos reportadas na literatura, o que restringiu o preenchimento da tabela. Além disso, alguns trabalhos não tiveram as bases mapeadas por não fornecerem as informações de interesse para essa análise e/ou uma descrição clara dos vídeos. A seguir, discutem-se limitações que dificultam ou impossibilitam a adoção das bases investigadas.

Um grupo de autores, identificados na Tabela 5, compõe a base a partir do site Wikio, ressaltando que o conteúdo engloba diversas categorias (documentário, notícias, música, propaganda, desenho animado, filme e esporte), sendo, portanto, genérico. Porém, o site apontado como referência não se encontra disponível para acesso e, em cada trabalho, seleciona-se um conjunto diferente de vídeos, dificultando comparações.

Outra opção empregada foi utilizar vídeos obtidos do YouTube por meio de consultas. Essa abordagem possibilita que se variem as buscas visando incluir diversas categorias, por exemplo, fenômenos naturais, shows aéreos, política, entretenimento, eventos sociais e esporte (WANG *et al.*, 2012). Em geral, os trabalhos apresentam as consultas utilizadas para a captura dos vídeos. Entretanto, a plataforma YouTube é muito dinâmica, conteúdos são constantemente adicionados e removidos. Logo, resultados retornados em determinado momento, podem não estar disponíveis algum tempo depois, dificultando a replicabilidade.

Aponta-se ainda que diversos conjuntos de vídeos possuem domínios específicos, uma restrição, muitas vezes, necessária para a avaliação, mas que dificulta a adoção da base por trabalhos de contextos diferentes. Exemplos incluem: vigilância; viagens, contendo atrações turísticas (Tour20 e Portal GeoVid); eventos cotidianos como trocar pneus, fazer lanches, cuidar de cachorros (TVSum); vídeos de usuários, reportando atrações turísticas e locais comuns como escolas, supermercados, piscina e parques.

Vale ressaltar que uma das técnicas analisadas concentra-se no domínio de notícias, assim como esta tese, mas trabalha com vídeos de um telejornal japonês, possuindo, portanto, a língua como principal barreira para sua utilização.

Tabela 5 – Principais características das bases de vídeo utilizadas para avaliação em trabalhos relacionados a sumarização multivídeo.

Nome	Domínio	Conjuntos	Vídeos por conjunto	Total de vídeos	Duração [v]: por vídeo [t]: total	Trabalhos
Wikio	Genérico	12	3-8	66		Wang e Merialdo (2009)
		63	3-8		1-7min [v]	Li e Merialdo (2010a)
		88	3-15			Li e Merialdo (2010b)
		30			0,5-10min [v]	Li e Merialdo (2011)
		21	3-15			Li <i>et al.</i> (2011)
		36	3-15	194	0,17-10min [v]	Li e Merialdo (2012)
		65	3-16			Li e Merialdo (2016)
YouTube	Genérico	12	Média: 85	1024	102h [t]	Hong <i>et al.</i> (2009)
		60		10130	751h [t]	Wang <i>et al.</i> (2012)
MVS1K	Genérico	10	83-109	936	31,5h [t]	Ji <i>et al.</i> (2018) Ji <i>et al.</i> (2019) Ji <i>et al.</i> (2020)
CoSum	Genérico	10		51		Panda e Roy-Chowdhury (2017)
TVSum50	Genérico	10		50		Panda e Roy-Chowdhury (2017)
NII TV-RECS	Notícias				3000h [t]	Ide <i>et al.</i> (2017)
	Vigilância	3	3,4,19	26		Ou <i>et al.</i> (2014) Ou <i>et al.</i> (2015)
		3	2,3,5	10		Zhu, Liao e Li (2016)
		4	2-5			Zhang <i>et al.</i> (2020)
Tour20	Viagens	20		140	7h [t]	Panda, Mithun e Roy-Chowdhury (2017) Wu, Zhong e Liu (2020)
Portal GeoVid	Viagens	4		72	3,3h [t]	Zhang <i>et al.</i> (2012) Zhang, Ma e Zimmermann (2013) Zhang, Zhang e Zimmermann (2015)
		9				Zhang e Zimmermann (2016)
	Vídeos de usuários	12	10-17	184		Zhang <i>et al.</i> (2015a)
TVSum	Eventos cotidianos	10	5		1-11min [v]	Wu, Zhong e Liu (2020)

Fonte: Dados da pesquisa.

3.4 Considerações finais

O levantamento realizado evidencia que os trabalhos relacionados a sumarização multivídeo fazem uso, para tarefas de avaliação, de uma diversidade de bases de vídeo. Porém, a maioria delas é limitada a contextos específicos, o que dificulta a reutilização. Em particular, para o domínio de notícias, não se identificou uma base aplicável.

A análise da literatura também permite verificar que diversos trabalhos adotam critérios

em comum para seleção de conteúdo embora, em geral, não utilizem a mesma combinação. Em vários casos, as técnicas empregam mais de um critério, visando aprimorar seus resultados e cobrir pontos de interesse para um domínio específico. Vale ressaltar que diferentes domínios são considerados nos trabalhos investigados, incluindo: notícias, vigilância, web (por exemplo, YouTube), *smartphones* e vídeos de usuários. Além disso, a estratégia utilizada para avaliar se um segmento de vídeo satisfaz os critérios varia para cada grupo de autores. Os métodos estudados propõem métricas específicas, como discutido anteriormente.

Portanto, pode-se concluir que a definição dos critérios não é endereçada em detalhes. Em geral, reporta-se o que dado critério significa para um domínio de vídeo em particular, isto é, qual informação acrescenta ao sumário final. Entretanto, não há discussões justificando como, entre todas as possibilidades, os autores decidiram que uma determinada combinação é, de fato, a melhor escolha. Além disso, nenhum dos trabalhos encontrados na literatura investiga o processo adotado por humanos para a produção de sumários. Tal estratégia possibilitaria mapear julgamentos de relevância aplicados por usuários e, como consequência, aproximar os sumários às suas expectativas.

CRITÉRIOS PARA SELEÇÃO DE CONTEÚDO: UM ESTUDO COM USUÁRIOS

Com o objetivo de investigar estratégias humanas empregadas no processo de sumarização multivídeo, foi desenvolvido um estudo com usuários finais. Da análise das estratégias, foi possível derivar critérios de seleção de conteúdo que podem aprimorar os resultados de técnicas automáticas.

Com métodos baseados no usuário é possível experimentar uma situação realística (PETRIE; BEVAN, 2009), obtendo a visão do público-alvo. De acordo com Sears e Hanson (2012), testes apropriados com usuários têm potencial para avançar no entendimento de problemas, levando ao desenvolvimento de soluções mais eficazes e melhorando a vida dos indivíduos estudados.

4.1 Descrição do estudo

Neste estudo, os usuários foram convidados a produzir, em duplas, um sumário a partir de um conjunto de vídeos relacionados, discutindo entre eles os motivos para inclusão ou exclusão de cada segmento de vídeo.

Para a atividade de sumarização, foram utilizados eventos pertencentes a diferentes categorias (esporte, tragédia, política, etc.). Cada evento envolve um conjunto de três vídeos de notícias, provenientes de telejornais brasileiros. Na Tabela 6, apresentam-se detalhes da base empregada no experimento. É possível observar que os conjuntos referem-se a tópicos de destaque na mídia nacional e possuem diferentes durações.

Tabela 6 – Descrição da base de vídeos utilizada no experimento.

Telejornal	Título da notícia	Data de exibição	Tempo vídeo*	Link
Olimpíadas 2016: Thiago Braz ouro no salto com vara				
Bom Dia Brasil	Thiago Braz é ouro no salto com vara e bate recorde olímpico	16/08/2016	6:39	Globo Play
Jornal Nacional	Público mostra seu carinho por Thiago Braz após o ouro no salto com vara	16/08/2016	3:35	Globo Play
Hora 1	Thiago Braz é medalha de ouro no salto com vara	16/08/2016	2:58	Globo Play
Queda do avião da Chapecoense				
Jornal Nacional	Avião com time da Chapecoense cai na Colômbia e mata 71 pessoas	29/11/2016	5:49	Globo Play
Jornal da Record	Queda de avião da Chapecoense deixa 71 mortos e mobiliza todo o país	29/11/2016	3:21	YouTube
Jornal da Band	Avião com integrantes da Chape cai e deixa 71 mortos	29/11/2016	4:56	Site Jornal
Posse Donald Trump				
Jornal Nacional	Trump assume a presidência dos EUA criticando antecessores	20/01/2017	7:59	Globo Play
Jornal da Record	Donald Trump toma posse em tradicional cerimônia em Washington (EUA)	20/01/2017	5:13	YouTube
Jornal da Band	Donald Trump toma posse nos Estados Unidos	20/01/2017	5:35	Site Jornal
Separação Angelina Jolie e Brad Pitt				
Jornal Hoje	Casal Angelina Jolie e Brad Pitt decidem se separar	20/09/2016	0:51	Globo Play
Jornal da Record	Vício e traição podem ter motivado a separação de Angelina Jolie e Brad Pitt	20/09/2016	1:52	YouTube
Jornal da Band	Depois de 12 anos casados, Angelina Jolie e Brad Pitt estão separados	20/09/2016	0:41	Site Jornal

*Tempo em minutos.

Fonte: Dados da pesquisa.

Antes da realização do experimento, cada usuário assinou o Termo de Consentimento Livre e Esclarecido (Apêndice A.1). Na Figura 13, apresenta-se uma visão geral do estudo, dividido em três etapas detalhadas a seguir.

Questionário sobre informações demográficas. Aplicou-se um questionário pré-sessão (Apêndice A.2) cujo objetivo foi obter informações demográficas e o interesse pelo tópico de pesquisa.

Sumarização multivídeo, em duplas, usando *Think Aloud*. Para a execução do experimento, fez-se o casamento aleatório entre evento e dupla de participantes, com a restrição de que cada conjunto de vídeos fosse sumarizado o mesmo número de vezes. Os participantes

Figura 13 – Visão geral do estudo realizado visando revelar estratégias humanas para sumarização multi-vídeo.

Fonte: Elaborada pelo autor.

foram informados de que o objetivo das tarefas era a sumarização, isto é, a criação de um único vídeo a partir do conjunto de entrada, de acordo com os critérios que considerassem mais relevantes.

Adotou-se o protocolo *Think Aloud*, usualmente empregado quando o foco é entender o comportamento do usuário, suas intenções e expectativas (PETRIE; BEVAN, 2009). Como o objetivo deste experimento era compreender como as pessoas produzem sumários multivídeo, ter participantes compartilhando seus pensamentos e explicando o que estão fazendo auxilia os pesquisadores a observar e traçar conclusões adequadas. Entretanto, momentos de silêncio durante o *Think Aloud* podem atrapalhar a obtenção de resultados. Assim, uma possível solução, discutida por Preece, Rogers e Sharp (2002), é propor a realização da atividade em dupla, possibilitando a conversa entre os participantes. Trabalhar com outra pessoa é mais natural e revelador, já que os membros da dupla irão interagir para se ajudarem. Esta etapa do estudo engloba quatro tarefas, descritas a seguir.

Treinamento na ferramenta de edição. Participar de um rápido treinamento sobre como realizar a sumarização de vídeos com uma ferramenta de edição. Neste estudo, utilizou-se o *software* Windows Movie Maker¹ devido a sua interface intuitiva.

Assistir ao conjunto de vídeos. Assistir ao conjunto de vídeos relacionados ao evento e compreendê-lo.

¹ Windows Movie Maker é uma ferramenta para criação e edição de vídeos. Possibilita a inclusão de efeitos, transições, títulos, créditos, áudio, etc.

Edição dos vídeos. Realizar a edição dos vídeos, isto é, o recorte de segmentos de interesse.

Produção do sumário. Produzir o sumário, discutindo com outro participante os motivos para inclusão ou exclusão de segmentos de vídeo. Tal passo foi gravado para posterior análise e anotações foram realizadas pela pesquisadora. Vale ressaltar que as anotações foram derivadas da observação dos usuários e continham informações relacionadas aos conteúdos que consideravam ou não relevantes, bem como as justificativas apresentadas.

Questionário sobre as tarefas de sumarização. Aplicou-se um questionário pós-sessão (Apêndice A.3) visando obter informações sobre a realização das tarefas e a relevância do tópico de pesquisa. Nesse questionário, os participantes podiam: 1) indicar se usaram ou não alguns critérios derivados de trabalhos relacionados apresentados em uma lista e 2) adicionar novos critérios que consideraram importantes.

Realizou-se um teste piloto com duas duplas para validar o planejamento do experimento e solucionar possíveis problemas. O teste revelou que alguns aprimoramentos pontuais poderiam ser feitos nos questionários e na condução do experimento: inclusão, no questionário pós-sessão, da alternativa que possibilita que o usuário informe um critério de seleção de conteúdo não coberto pelas opções de resposta (campo “Outro”); necessidade de se explicar detalhadamente aos participantes o que se espera como resposta, para se obter informações mais específicas e adequadas às perguntas. Porém, de maneira geral, o teste piloto mostrou que o planejamento estava alinhado aos resultados esperados e era funcional, pois os usuários foram capazes de compreender os objetivos do estudo e realizar as tarefas sem dificuldades.

A seleção dos participantes ocorreu por meio de um convite realizado pessoalmente ou por e-mail a indivíduos com o seguinte perfil: 1) escolaridade mínima: ensino médio; 2) nível mínimo de conhecimentos em informática: básico. A definição do perfil englobou habilidades consideradas suficientes para os participantes interagirem com uma interface de edição de vídeos, bem como assistirem e interpretarem o conteúdo dos mesmos. O processo de recrutamento contou com o altruísmo das pessoas em ajudar um trabalho científico e sua curiosidade e interesse pelo tópico de pesquisa.

O estudo envolveu 24 participantes, divididos em 12 duplas. Nas Figuras 14a e 14b, pode-se verificar que os usuários possuem idades entre 23 e 54 anos, sendo 67% do sexo masculino. Já nas Figuras 14c e 14d, é possível observar a distribuição entre os níveis de escolaridade e conhecimentos em informática, respectivamente. Por fim, na Figura 14e, indicam-se as profissões dos participantes, com a respectiva distribuição apresentada entre parênteses. Os dados demográficos mostram a diversidade de perfis dos envolvidos no experimento.

Para a execução do estudo, os voluntários foram divididos em duplas e contatados via ligação telefônica ou mensagem de texto para agendamento de horários convenientes para os

Figura 14 – Dados demográficos dos participantes do experimento.

Fonte: Dados da pesquisa.

envolvidos. As sessões presenciais duraram, em média, 59 minutos e aconteceram em um período de 18 dias.

4.2 Análise qualitativa

Nesta seção, apresenta-se a análise realizada com o objetivo de extrair os dados de interesse do experimento, isto é, compreender como os usuários realizam o processo de sumarização multivídeo e derivar critérios para seleção de conteúdo que possam guiar o desenvolvimento de técnicas automáticas. Os critérios identificados são definidos e discutidos no contexto do estudo.

Com base na observação da produção dos sumários pelos participantes do experimento e também nas respostas ao questionário pós-sessão, foram obtidas estratégias humanas empregadas no processo de sumarização multivídeo. Na Tabela 7, apresentam-se os critérios de seleção de conteúdo aplicados pelos usuários. Para extrair tais informações, foram considerados critérios

utilizados por, pelo menos, duas duplas na respectiva coleção de vídeos, ou seja, que representavam consenso entre os participantes. Porém, não se impôs a restrição de que as estratégias identificadas fossem comuns a todos os conjuntos de vídeos, pois algumas delas são reveladas em contextos específicos.

Tabela 7 – Estratégias humanas para sumarização multivídeo identificadas no estudo com usuários.

Critérios de inclusão
Introdução
Cronologia
Fluência
Redundância
Informação única
Cobertura
Conteúdo visual
Segmentos com fonte de informação
Critérios de exclusão
Subjetividade
Informações complementares
Redundância
Comentários ou opiniões de jornalistas
Entrevistas

Fonte: Dados da pesquisa.

Primeiramente, são introduzidas definições para os critérios de seleção de conteúdo presentes na Tabela 7, fornecendo uma visão geral das descobertas. Então, apresenta-se uma discussão mais detalhada com exemplos obtidos do estudo com usuários.

Critérios de inclusão

Introdução. Selecionar um segmento que apresenta o tópico principal abordado pela notícia, fornecendo uma visão geral de seu conteúdo.

Cronologia. Segmentos de diferentes vídeos são inseridos no sumário considerando a ordem cronológica, gerando um resultado mais compreensível com uma sequência coerente.

Fluência. As transições entre os segmentos selecionados devem ser suaves, considerando aspectos visuais, sonoros e textuais, de modo que não prejudiquem a experiência de visualização do usuário.

Redundância. A importância de um segmento é avaliada por repetição de conteúdo, ou seja, uma heurística comum utilizada pelos usuários é inferir que uma informação que aparece em segmentos de vídeos distintos é relevante.

Informação única. Um segmento é selecionado caso seu conteúdo apareça em apenas um vídeo de entrada e seja fortemente relacionado ao tópico principal da notícia.

Cobertura. Um segmento é incluído no sumário se é capaz de tornar o resultado mais completo, cobrindo o máximo possível de informações relevantes.

Conteúdo visual. Um segmento é escolhido com base em seu conteúdo visual, isto é, se contém imagens de interesse.

Segmentos com fonte de informação. Um segmento é considerado relevante se apresenta a fonte da informação discutida.

Crítérios de exclusão

Subjetividade. Um segmento não é levado ao sumário se não apresenta a informação de modo objetivo mas, ao contrário, usa conteúdo emocional ou emprega muitos adjetivos.

Informações complementares. São excluídos segmentos com conteúdo que não está diretamente relacionado ao tópico principal, como curiosidades ou fatos secundários.

Redundância. Um segmento não é selecionado para compor o sumário caso contenha conteúdo similar ao dos segmentos já incluídos, aumentando a diversidade de informações apresentadas ao usuário.

Comentários ou opiniões de jornalistas. Segmentos que contêm comentários ou opiniões de jornalistas não são considerados relevantes.

Entrevistas. Não são levados ao sumário segmentos mostrando entrevistas.

O estudo revelou, conforme apontado na Tabela 7, que uma das estratégias empregadas pelos usuários foi iniciar o sumário com uma **introdução** pertencente a um dos vídeos do conjunto. Tal introdução é composta pela apresentação do evento pelos âncoras do telejornal, representando um resumo do fato ou, em analogia às mídias impressas, o título da notícia. Segundo os usuários, essa informação é importante por fornecer uma contextualização do que será apresentado em seguida e, conseqüentemente, facilitar o entendimento do conteúdo.

Os usuários também destacaram o uso da **cronologia**. Esse critério foi considerado particularmente importante para os tópicos em que havia uma história a ser contada, logo, ordenar os acontecimentos possibilitava a criação de uma narrativa consistente e compreensível. Por exemplo, nos vídeos a respeito da posse de Donald Trump, a estratégia empregada foi descrever o dia do presidente e, por isso, a ordem dos eventos era relevante. Outro exemplo é o conjunto relacionado às Olimpíadas, em que se julgou essencial relatar, de maneira ordenada, os fatos até a obtenção dos resultados da disputa.

Não somente a cronologia mostrou-se um ponto pertinente, como também a **fluência** do sumário. Os usuários indicaram ser importante que as transições de segmentos encaixem de maneira a fazer sentido e proporcionar continuidade. Foi possível notar essa preocupação nos participantes, durante a realização do experimento, ao posicionarem um segmento selecionado

em determinado ponto do sumário e executarem uma prévia do vídeo resultante, discutindo entre si se a transição estava suave o suficiente.

Além disso, o estudo confirmou que **redundância**, isto é, informações que se repetem em diversos vídeos do conjunto demonstram relevância e são usualmente inseridas no sumário. Exemplos incluem: primeiro discurso de Trump como presidente; eventos tradicionais do dia da posse; disputa entre o atleta brasileiro e o favorito francês, resultando na quebra do recorde olímpico do salto com vara; número de mortes no acidente de avião.

Embora, a princípio, possa parecer uma estratégia contraditória à anteriormente discutida, **informações únicas** também foram consideradas na composição do vídeo resultante. Esse critério está relacionado ao de cobertura. Os usuários salientaram que, se um dado é apresentado em apenas uma versão da notícia, ele é importante para complementar o conteúdo. Por exemplo, no conjunto de vídeos que aborda a queda do avião da Chapecoense, apenas o Jornal da Band noticiou a identidade dos sobreviventes da tragédia. Para os participantes do experimento, não era suficiente saber que seis pessoas conseguiram se salvar, eles gostariam de ser informados sobre quem eram esses jogadores, jornalistas e tripulantes. Outros exemplos foram observados durante a realização do experimento, como a inclusão no sumário da duração do discurso de Donald Trump e da altura do salto realizado pelo competidor francês em uma tentativa de vencer o atleta brasileiro.

A **cobertura** em si foi mais um fator de interesse. Os usuários procuraram relatar o fato principal com o máximo possível de informações, incluindo os decretos assinados pelo presidente em seu primeiro dia de exercício, o tradicional desfile, guarda dos filhos do casal de atores após o divórcio, os motivos que levaram à separação, o que o casal falou a respeito, investigações do acidente com o avião da Chapecoense, informações sobre o resgate e sobreviventes.

Figura 15 – Exemplos do critério de inclusão conteúdo visual.

(a) Quadro de vídeo mostrando as alturas já ultrapassadas pelo atleta brasileiro. (b) Quadro de vídeo apresentando a rota do avião da Chapecoense.

Fonte: Elaborada pelo autor.

Adicionalmente, o **conteúdo visual** foi considerado pelos usuários. Foram selecionados quadros contendo objetos de interesse, como as alturas já ultrapassadas pelos competidores na

prova de salto com vara (Figura 15a), o novo recorde olímpico e documentos comprovando o pedido de divórcio. Além disso, segmentos com animações ou simulações foram incluídos no sumário (Figura 15b). Essas demonstrações, em geral, estão relacionadas a explicações que podem facilitar o entendimento da notícia. Segundo os participantes, o conteúdo visual é importante para enriquecer os resultados e auxiliar a tornar o sumário mais compacto, uma vez que algumas informações podem ser apresentadas nos quadros, complementando ou substituindo o áudio.

Os participantes do experimento deram preferência a segmentos que citavam a **fonte da informação**, por proporcionar credibilidade. Por exemplo, na coleção de vídeos sobre a separação de Angelina Jolie e Brad Pitt, mencionava-se uma nota do advogado da atriz ou o *site* TMZ.

O estudo também permitiu revelar critérios utilizados pelos usuários para excluir informações. O primeiro deles foi eliminar **subjetividade**, isto é, segmentos de vídeo que continham muitos adjetivos e não retratavam a notícia de maneira objetiva, como a descrição do casal de atores como “ricos”, “engajados”, “símbolo da felicidade”, “casal perfeito e emblemático” ou ainda referências à carreira do atleta olímpico brasileiro, ressaltando que era conhecido como “covarde” e “amarelão”. Na visão dos participantes do experimento, tais caracterizações soam sensacionalistas, exploram os sentimentos das pessoas e não acrescentam informações sobre o evento em si. Por esses motivos, não deveriam ser levadas ao sumário.

Além disso, os usuários não julgaram relevante selecionar **informações complementares**, como fatos secundários e curiosidades. Como o objetivo de sumário multivídeo é prover uma versão reduzida do conteúdo apresentado em um conjunto de vídeos, os participantes do experimento consideraram importante concentrar-se no tópico principal, incluindo informações diretamente relacionadas a ele. Exemplos de pontos excluídos incluem: histórico do atletismo, especulações sobre um filme baseado na vida de Angelina Jolie e Brad Pitt, o discurso de despedida do presidente Obama. Vale ressaltar que o critério de inclusão “cobertura” está vinculado ao foco da notícia, para os usuários é importante que o assunto principal seja retratado de maneira completa e não que tudo que é abordado nos vídeos seja adicionado.

Redundância foi outro critério de exclusão empregado. Essa estratégia também foi utilizada pelos usuários para inclusão de conteúdo, porém o objetivo, nesse caso, é evitar repetições. Quando várias fontes apresentavam a mesma informação, apenas uma versão era escolhida para compor o sumário.

Por fim, **comentários ou opiniões de jornalistas**, bem como **entrevistas** foram excluídos pelos usuários. Isso pode ser explicado pelo fato de que tais segmentos representam visões particulares e parciais sobre o assunto, muitas vezes refletindo a filosofia da emissora ou conceitos e valores do indivíduo.

O estudo desenvolvido revelou critérios utilizados pelos usuários para selecionar conteúdo a ser incluído no sumário e para filtrar informações consideradas menos relevantes. Foi possível verificar que os critérios estão relacionados principalmente a aspectos semânticos e de apresentação e, alguns deles, mostram-se mais importantes em contextos específicos. Um exemplo desse último caso é o critério segmentos com fonte de informação, já que se trata de um conteúdo que nem sempre está presente nos vídeos.

4.3 Análise quantitativa

Nesta seção, apresentam-se análises dos dados obtidos no estudo com o objetivo de: 1) investigar a relevância de cada critério de seleção de conteúdo, considerando-se seu uso pelos usuários durante o experimento; 2) realizar uma comparação com a literatura, isto é, com critérios adotados por trabalhos relacionados a sumarização multivídeo.

Para o primeiro ponto de interesse, na Tabela 8, é possível observar a porcentagem de participantes que utilizaram cada critério no processo de desenvolvimento dos sumários. Os valores foram obtidos de duas maneiras: 1) observação e questionário - os dados foram derivados da observação dos participantes durante o experimento e suas respostas ao questionário pós-sessão; 2) análise dos sumários - os sumários gerados foram investigados com o intuito de verificar se os usuários realmente aplicaram os critérios que apontaram como relevantes.

Tabela 8 – Porcentagem de participantes que aplicaram cada critério. Dados obtidos pela: 1) observação dos usuários durante o experimento e suas respostas ao questionário pós-sessão e 2) investigação dos sumários produzidos.

	Observação e questionário (%)	Análise dos sumários (%)
Inclusão		
Introdução	92	92
Cronologia	67	89
Fluência	75	100
Redundância	92	100
Informação única	59	84
Cobertura	75	84
Conteúdo visual	89	78
Segmentos com fonte de informação	50	100
Exclusão		
Subjetividade	89	78
Informações complementares	84	58
Redundância	100	92
Comentários ou opiniões de jornalistas	100	89
Entrevistas	78	78

Fonte: Dados da pesquisa.

Analisando-se a Tabela 8, é possível contrastar o que os usuários disseram com o que realmente fizeram. Desse modo, pode-se observar uma convergência nos valores para os critérios de inclusão: introdução, redundância, cobertura e conteúdo visual e critérios de exclusão: subjetividade, redundância, comentários ou opiniões de jornalistas e entrevistas. Nesses casos, a diferença entre as abordagens é em torno de 10% ou menos, mostrando que os participantes expressaram suas preferências e, de fato, as utilizaram para produzir os sumários.

Por outro lado, pode-se verificar que cronologia, fluência, informação única e segmentos com fonte de informação mostraram-se mais relevantes quando os sumários foram analisados do que havia sido inicialmente inferido somente pelo protocolo *Think Aloud* e questionários. Isso pode ser explicado pelo fato de que esse estudo foi conduzido visando descobrir critérios para a sumarização multivídeo, não havia uma lista padronizada previamente disponível a todos. Assim, é possível que alguns usuários não tenham externalizado um critério durante o experimento, mas o utilizaram no processo de desenvolver um sumário. Tais casos confirmam que a investigação dos resultados gerados é essencial para complementar as análises e conclusões.

É importante ressaltar que cobertura, redundância e informação única são critérios relacionados. Pode-se pensar na cobertura como um critério derivado de redundância e informação única. Em outras palavras, cobertura significa que os participantes têm a preocupação de incluir a maioria do conteúdo associado ao assunto principal, o que pode ser feito selecionando-se segmentos que aparecem em diversos vídeos (redundância) e em vídeos específicos (informação única). Isso pode justificar a diferença nos valores obtidos: alguns usuários apenas indicaram cobertura, apesar de também terem usado informação única. Por exemplo, para o conjunto de vídeos relacionados à queda do avião da Chapecoense, apenas o Jornal Nacional noticiou que as caixas pretas da aeronave já haviam sido encontradas pelos investigadores e os dados e as gravações das conversas na cabine iriam ajudar a esclarecer as causas do acidente. Nesse caso, houve participantes do experimento que incluíram esse conteúdo no sumário, mas não citaram informação única como critério utilizado, apontando apenas que se preocuparam em deixar o resultado completo (cobertura).

Para informações complementares, pode-se observar a situação oposta, isto é, os participantes aplicaram esse critério menos do que o declararam como importante. Tal caso reflete disparidades no conceito de informações complementares, sendo que segmentos que alguns usuários consideram relacionados ao tópico principal podem ser classificados como curiosidades ou fatos secundários. Por exemplo, no conjunto de vídeos sobre a vitória de Thiago Braz na prova de salto com vara, alguns participantes do experimento julgaram que a história do atleta brasileiro era relevante ao contexto, incluindo que possuía 22 anos, era o campeão mundial júnior e foi criado pela avó. Entretanto, é possível considerar que apenas o conteúdo relacionado à competição que levou ao ouro deveria compor o sumário, sendo as informações pessoais rotuladas como curiosidades/fatos secundários.

Além disso, pode-se comparar os critérios identificados neste estudo com usuários (Tabela 8) com trabalhos relacionados reportados na literatura e discutidos no Capítulo 3 (Tabela 3). Para deixar as similaridades e diferenças mais evidentes, introduz-se a Tabela 9, que traz a compilação dos critérios de seleção de conteúdo para sumarização multivídeo derivados de estratégias humanas e presentes na literatura. Utiliza-se o sinal “v” para indicar que o critério compõe o conjunto e “x”, caso contrário.

Tabela 9 – Comparação entre os critérios revelados pelo estudo com usuários e os critérios identificados em trabalhos relacionados.

	Estratégias humanas	Literatura
Inclusão		
Introdução	✓	✗
Cronologia	✓	✓
Fluência	✓	✓
Redundância	✓	✓
Informação única	✓	✗
Cobertura	✓	✓
Conteúdo visual	✓	✗
Segmentos com fonte de informação	✓	✗
Qualidade/estética	✗	✓
Objetos em movimento	✗	✓
Exclusão		
Subjetividade	✓	✗
Informações complementares	✓	✗
Redundância	✓	✓
Comentários ou opiniões de jornalistas	✓	✗
Entrevistas	✓	✗

Fonte: Dados da pesquisa.

Primeiramente, é importante apontar alguns critérios que aparecem em trabalhos relacionados e no estudo com usuários com terminologias diferentes, mas possuem significado semelhante. Pode-se mapear o critério **diversidade** da Tabela 3 com o critério de exclusão **redundância** da Tabela 8, pois compartilham o objetivo de evitar informações repetidas. Também **consistência visual** dos trabalhos relacionados pode ser considerada similar a **fluência** derivada de estratégias humanas. Entretanto, vale mencionar que fluência é um critério mais amplo e cobre não apenas os aspectos visuais, mas toda a experiência de assistir a um sumário, incluindo elementos de áudio e/ou texto. Nas discussões a seguir adota-se a terminologia utilizada no estudo.

Assim, é possível encontrar sobreposições entre os critérios destacados pelos usuários e os que aparecem na literatura: 1) inclusão - cronologia, fluência, redundância, cobertura; 2) exclusão - redundância.

Redundância é considerado o critério de seleção de conteúdo mais importante tanto para

o caso de inclusão, quanto para exclusão, havendo uma diferença na medida de relevância. É possível notar alta convergência nos resultados baseados em estratégias humanas, enquanto os trabalhos encontrados na literatura apresentam uso de 58,6% (Tabela 3). Cronologia, fluência e cobertura, por outro lado, são considerados importantes na visão do usuário, mas pouco explorados por trabalhos relacionados (6,9%, 17,2% e 13,8%, respectivamente - Tabela 3).

Além disso, é possível verificar que diversos critérios apontados pelos usuários não aparecem na análise da literatura: 1) inclusão - introdução, informação única, conteúdo visual, segmentos com fonte de informação; 2) exclusão - subjetividade, informações complementares, comentários ou opiniões de jornalistas, entrevistas. Na situação oposta, dois critérios empregados por técnicas reportadas na literatura não são relevantes de acordo com os usuários. Um deles é objetos em movimento, que é um critério muito específico aplicado no domínio de vigilância. O outro é estética, indicando que as pessoas estão mais interessadas no conteúdo abordado em um segmento do que em seus aspectos de qualidade visual.

Portanto, pode-se concluir que os sumários produzidos pelos trabalhos relacionados não estão totalmente alinhados com as expectativas dos usuários. Também é importante destacar que a maioria das técnicas de sumarização multivídeo são baseadas em características de baixo nível, enquanto os critérios derivados de estratégias humanas consideram semântica, isto é, o entendimento do conteúdo do vídeo e, por essa razão, apresentam desafios para o desenvolvimento.

4.4 Considerações finais

Neste capítulo, apresentou-se um estudo com usuários desenvolvido para revelar estratégias humanas para sumarização multivídeo, aplicado ao domínio de notícias. Discutiu-se a metodologia do estudo, bem como os critérios dele derivados.

A análise dos resultados evidenciou que alguns critérios reportados na literatura também são importantes na visão do usuário (cronologia, fluência, redundância, cobertura e diversidade) e apontou abordagens diferentes para a produção de sumários de vídeos de notícias, ainda não exploradas por trabalhos relacionados, que são: 1) critérios de inclusão - introdução, informação única, conteúdo visual, segmentos com fonte de informação; 2) critérios de exclusão - subjetividade, informações complementares, comentários ou opiniões de jornalistas, entrevistas.

No Capítulo 5, apresenta-se uma técnica automática de sumarização multivídeo que incorpora critérios de seleção de conteúdo derivados de estratégias humanas.

TÉCNICA DE SUMARIZAÇÃO MULTIVÍDEO

HS-MVIDEOSUMM

Neste capítulo, apresenta-se a técnica de sumarização multivídeo *Human Strategies based Multi-Video Summarization* (HS-MVideoSumm), que incorpora critérios para seleção de conteúdo baseados em estratégias humanas e identificados no estudo com usuários discutido no Capítulo 4. Para um melhor entendimento, este capítulo está organizado da seguinte maneira: primeiramente se apresenta uma visão geral da técnica, então, cada um de seus módulos é detalhado (Seções 5.1 a 5.5) e, por fim, resumem-se quais critérios revelados pelo estudo foram considerados no desenvolvimento da HS-MVideoSumm (Seção 5.6).

Figura 16 – Visão geral da HS-MVideoSumm, técnica de sumarização multivídeo baseada em estratégias humanas.

Fonte: Elaborada pelo autor.

A Figura 16 traz uma visão geral da técnica, ilustrando seus 5 módulos: Seleção da introdução, Detecção de subjetividade, Detecção de segmentos redundantes, Seleção de segmentos e Produção do sumário. A entrada esperada consiste em um conjunto de vídeos de notícias relacionadas a um evento e seus respectivos conteúdos textuais segmentados em subtópicos. Vale ressaltar que um subtópico é definido como um agrupamento de frases consecutivas abordando um mesmo assunto, inserido no tópico principal. A partir de então os módulos se sucedem,

linearmente, sendo que a saída de um constitui a entrada para o próximo e cada módulo incorpora um ou mais critérios de seleção apresentados na Tabela 7. Por fim, o resultado da técnica é um vídeo que representa uma versão compacta e informativa do conteúdo original, incluindo os segmentos considerados mais relevantes pelos processamentos realizados. No contexto deste trabalho, um segmento de vídeo é delimitado pelo tamanho do subtópico, contendo além do material textual, os dados visuais e sonoros associados.

Na Figura 17 apresenta-se a estrutura de um segmento processado pela técnica. É possível observar que é formado pelo conteúdo textual e marcações temporais de início e fim no vídeo. Tais informações fornecem uma referência para a obtenção de outros dados que possam ser necessários, por exemplo, os quadros de vídeo. Logo, a técnica HS-MVideoSumm recebe como entrada um conjunto de segmentos com essa estruturação, os quais vão sendo filtrados/selecionados em cada módulo, até a geração do sumário. Detalhes sobre o conteúdo textual utilizado neste trabalho são fornecidos na Seção 6.1.

Figura 17 – Exemplo de segmento processado pela técnica HS-MVideoSumm.

```
{  
  "content": "A parte da catedral mais castigada pelas chamas foi o teto. Nesse ponto,  
o local estava cercado de andaimes. O telhado gótico, construído na época medieval,  
passava por obras de restauração desde o ano passado. A polícia disse que a estrutura  
principal foi salva. As estátuas de bronze foram removidas na semana passada para as  
obras.",  
  "begin": "00:01:51",  
  "end": "00:02:13"  
},
```

Fonte: Dados da pesquisa.

A implementação foi desenvolvida em Python 3, com o auxílio de bibliotecas para manipulação de imagens e extração de características: *Open Source Computer Vision Library* (OpenCV)¹; aprendizado de máquina: *sklearn.cluster*²; edição de vídeos: *moviepy.editor*³.

5.1 Seleção da introdução

Este módulo engloba três etapas principais: 1) detecção das introduções de cada vídeo do conjunto de entrada; 2) escolha de uma introdução para ser levada ao sumário; 3) exclusão das introduções não selecionadas. A etapa 3 é considerada trivial por consistir na remoção dos segmentos mencionados dos dados carregados em memória, para que não sejam processados pelos módulos seguintes. As etapas 1 e 2, por sua vez, são detalhadas a seguir.

¹ OpenCV é uma biblioteca que provê recursos para aplicações de visão computacional e aprendizado de máquina (<https://opencv.org>).

² *sklearn.cluster* é uma biblioteca que reúne os algoritmos mais populares de agrupamento não supervisionado (<https://scikit-learn.org/stable/modules/classes.html#module-sklearn.cluster>).

³ *moviepy.editor* é uma biblioteca para edição de vídeos: recorte, concatenação, inserção de título, composição, processamento e criação de efeitos customizados (<https://pypi.org/project/moviepy/>).

Detecção das introduções

Baseia-se em características do domínio de vídeos de notícias, em que a introdução está localizada no início e é composta por um ou mais âncoras apresentando o tópico abordado, em geral, em estúdio. Assim, a transição de quadro que marca o fim da introdução e início da reportagem pode ser identificada pelo conteúdo visual. Como mostrado na Figura 18a, os quadros pertencentes à introdução são similares entre si e possuem elementos específicos, como o cenário e os âncoras, podendo ser visualmente diferenciados dos quadros que compõem o segmento posterior, exemplificado na Figura 18b.

Figura 18 – Exemplos de quadros pertencentes à introdução e segmento subsequente.

(a) Quadro de vídeo extraído da introdução.

(b) Quadro de vídeo extraído do segmento posterior à introdução.

Fonte: Elaborada pelo autor.

Na abordagem adotada, utiliza-se uma taxa de amostragem de um quadro por segundo, a partir do início do vídeo. Tal valor é considerado adequado para o domínio, uma vez que quadros sucessivos possuem conteúdo similar, não sendo observadas variações significativas em intervalos menores que um segundo. Por outro lado, a redução de informações apresenta vantagens quanto ao tempo de processamento.

Os quadros amostrados são representados por meio de histogramas em escala de cinza (Seção 2.2.2). Então, quando a medida de intersecção de histogramas de quadros consecutivos é menor que 70%, identifica-se o fim da introdução. Essa medida se destaca por ser robusta a variações no plano de fundo, visualização de objetos a partir de diferentes pontos de vista, oclusão e alterações na resolução das imagens (SWAIN; BALLARD, 1991). O limiar foi definido empiricamente na base de vídeos (Tabela 11), variando-se seu valor de 0.5 a 0.9, passo 0.05 e avaliando-se os resultados obtidos. Para obtenção dos quadros, empregou-se o *framework* FFmpeg⁴ e, para leitura das imagens, cálculo de histogramas e intersecção, utilizaram-se funções da biblioteca OpenCV.

Como a unidade básica processada pela técnica HS-MVideoSumm é o subtópico, pode ocorrer do quadro detectado como fim da introdução não corresponder ao limite de um

⁴ FFmpeg é um *framework* multiplataforma para gravar, converter e transmitir conteúdo de áudio e vídeo (<https://ffmpeg.org>).

segmento. Nesse caso, consideram-se como segmentos que integram a introdução aqueles que possuem apenas conteúdo de apresentação do tópico, ou seja, se um segmento é formado por trechos contendo os âncoras agrupados com parte da reportagem, ele não é considerado como introdução. Tal estratégia é adotada para que não se percam informações potencialmente relevantes, já que apenas uma opção de introdução será levada ao sumário e as demais serão excluídas.

Escolha da introdução

A menor introdução entre as detectadas no conjunto de vídeos de entrada é selecionada para ser posteriormente incluída no sumário. Esse critério foi adotado considerando que o conteúdo da introdução é importante para contextualizar o acontecimento retratado pela notícia, contribuindo para uma narração coerente. Entretanto, como o sumário tem o objetivo de ser o mais compacto possível, optou-se por uma breve apresentação, priorizando-se as informações do evento.

Portanto, após a aplicação deste módulo, tem-se como resultado uma introdução selecionada para compor o sumário e as demais excluídas de processamentos posteriores.

5.2 Detecção de subjetividade

Este módulo consiste em encontrar e excluir segmentos com informações subjetivas. Em vídeos de notícias, a subjetividade aparece em comentários ou opiniões de jornalistas e em entrevistas. Por incorporar critérios de exclusão, este módulo é aplicado em uma fase inicial da técnica, visando filtrar informações consideradas irrelevantes para o contexto e contribuindo para a redução do volume de dados para processamento em etapas posteriores.

Figura 19 – Visão geral do módulo para detecção de segmentos subjetivos.

Fonte: Elaborada pelo autor.

Na Figura 19, apresenta-se uma visão geral da solução desenvolvida para a detecção de subjetividade, composta de duas etapas. Na primeira etapa, aplicam-se técnicas de reconhecimento facial como um modo de verificar se determinado segmento possui indícios de opiniões ou entrevistas. Na segunda etapa, analisa-se sua informação textual para confirmar a presença de

subjetividade. Isso é feito por meio de uma abordagem que combina análise de sentimentos e detecção de adjetivos.

Considerando-se o domínio de vídeos de notícias, pode-se observar um padrão quando ocorrem opiniões de jornalistas ou entrevistas. Tais segmentos se concentram na pessoa que está falando. Na Figura 20, apresentam-se exemplos desses casos. É possível ver o rosto de um jornalista (Figura 20a) e de um entrevistado (Figura 20b) em destaque no quadro.

Figura 20 – Exemplos de quadros com faces em destaque.

(a) Segmento com opiniões de jornalista.

(b) Segmento com entrevista.

Fonte: Elaborada pelo autor.

Assim, o primeiro passo na busca por subjetividade em um segmento de vídeo é a detecção de faces. Para realizar essa tarefa, foram adotados recursos da biblioteca OpenCV. Utilizou-se um classificador *Local Binary Pattern* (LBP) pré-treinado para reconhecimento de faces⁵. Para cada segmento de vídeo, processando um quadro por segundo, verificou-se se havia, pelo menos, uma imagem com faces em destaque. Em caso positivo, o segmento foi marcado como candidato à remoção.

Entretanto, nem todos os segmentos com foco no narrador estão reportando opiniões. Para confirmar se a subjetividade está presente, foram analisadas as informações textuais relacionadas. Vale mencionar que comentários ou opiniões de jornalistas e entrevistas contêm conteúdo similar, isto é, pessoas oferecendo suas visões particulares sobre determinado assunto. Por isso, em um segundo passo, o objetivo é inspecionar se o texto presente em um segmento de vídeo é neutro, retratando o tópico de maneira objetiva ou apresenta polaridade, indicando uma opinião.

Para lidar com esse problema, propõe-se uma abordagem que combina: 1) análise de sentimentos e 2) detecção de adjetivos, ilustrada na Figura 21 e detalhada a seguir.

Análise de sentimentos

Usada para capturar a emoção em um segmento, determina a atitude de modo geral, positiva ou negativa, expressa em um texto.

⁵ https://docs.opencv.org/3.4.2/db/d28/tutorial_cascade_classifier.html

Adotou-se a *Google Cloud Natural Language API*⁶. Nessa plataforma, os sentimentos são representados por valores numéricos: *score* e magnitude. O *score* varia de -1 (negativo) a 1 (positivo) e corresponde à inclinação emocional de um texto. A magnitude indica a força da emoção e seu valor é, em geral, proporcional ao tamanho do documento, variando de 0 a +inf (infinito). Ambos os dados podem ser obtidos para um texto completo e/ou para cada uma de suas sentenças.

A plataforma aponta que os limiares aplicados para classificar um conteúdo como neutro ou positivo/negativo devem ser ajustados ao contexto de uso. Por isso, foram definidos três tamanhos para os segmentos: pequeno, médio e grande, com respectivos valores de magnitude considerados significativos para indicar a polaridade do texto. Para segmentos grandes, como o tamanho máximo pode variar, uma análise adicional é realizada para confirmar a polaridade: além de um valor mínimo de magnitude, o texto deve satisfazer a condição de que, pelo menos, 40% de suas sentenças tenha *score* não neutro. Isso indica que uma parte relevante do segmento é composta por sentenças com conteúdo emocional.

Detecção de adjetivos

Essa solução provém dos usuários. No estudo realizado (Capítulo 4), os usuários finais argumentaram que a subjetividade está relacionada ao emprego de vários adjetivos. Assim, definiu-se um número de adjetivos para cada tamanho de segmento (pequeno, médio e grande), indicando que a informação não é apresentada de maneira objetiva. Para isso, utilizou-se o SentiLex-PT⁷, um léxico para português, do qual é possível obter a classificação das palavras.

É importante destacar que os limiares apresentados na Figura 21 foram definidos empiricamente para o contexto de vídeos de notícias. Considera-se que a base utilizada nos experimentos (Tabela 11) é representativa do domínio, uma vez que contém vídeos de diferentes durações e categorias. Dessa maneira, seus segmentos foram investigados e verificou-se que 42% possuía tamanho menor que 35 palavras, 42% entre 35 e 70 palavras, restando uma pequena porcentagem (16%) maior que 70 palavras. Tal análise guiou a definição dos intervalos considerados “pequeno”, “médio” ou “grande”. A porcentagem de segmentos na categoria “grande” mostra que esse é um caso de exceção e, devido a impossibilidade de se fixar um tamanho máximo, uma condição extra é incluída, como já explicado. Então, os valores de magnitude, *score* e número de adjetivos foram ajustados para cada caso, variando-se os dados e avaliando-se os resultados da classificação do segmento (subjetivo ou neutro).

⁶ A *Natural Language API* possui modelos pré-treinados que permitem aos desenvolvedores trabalhar com recursos de processamento de linguagem natural, incluindo análise de sentimento, análise de entidades, análise de sentimento de entidades, classificação de conteúdo e análise sintática (<https://cloud.google.com/natural-language/docs/basics>). Neste trabalho, foi utilizada em janeiro/2020.

⁷ SentiLex-PT é um léxico para português composto por 7.014 lemas e 82.347 formas flexionadas. É especialmente útil para aplicações de mineração de opinião (<http://b2find.eudat.eu/dataset/b6bd16c2-a8ab-598f-be41-1e7aeecd60d3>).

Figura 21 – Critérios para classificar um segmento textual como subjetivo: opinião de jornalista ou entrevista.

Fonte: Elaborada pelo autor.

Os segmentos rotulados como subjetivos são, então, excluídos de processamentos posteriores. Logo, este módulo atua como um filtro de informações, incorporando os critérios de exclusão relacionados a subjetividade: comentários ou opiniões de jornalistas e entrevistas.

5.3 Detecção de segmentos redundantes

Neste módulo, o critério de inclusão redundância foi integrado à técnica, isto é, foram identificadas informações que se repetem no conjunto de vídeos de entrada. Para isso, a abordagem proposta baseia-se no conteúdo textual dos segmentos e pode ser dividida em três etapas principais, ilustradas na Figura 22 e detalhadas a seguir: 1) pré-processamento; 2) *Bag of Words*; 3) busca por correspondências.

Figura 22 – Visão geral do módulo para detecção de segmentos redundantes.

Fonte: Elaborada pelo autor.

Primeiramente, os textos dos segmentos passam por um **pré-processamento**, em que é feita a conversão dos caracteres para minúsculo, a remoção de sinais de pontuação, números e

stopwords, bem como a redução de cada palavra ao seu radical (do inglês, *stemming*). Para a implementação foram utilizados recursos do *Natural Language Toolkit* (NLTK)⁸. Essa plataforma contém um dicionário com as palavras consideradas *stopwords* na língua portuguesa e funções para o processo de *stemming*.

Então, o texto pré-processado é representado por meio do modelo **Bag of Words** (Seção 2.2.1): obtém-se o vocabulário da coleção de segmentos e cada um deles é descrito por um vetor de características. Para cada posição do vetor, calcula-se o valor TF-IDF (Seção 2.2.1), que utiliza o número de ocorrências de um termo no segmento, ponderado pela sua frequência na coleção. O intuito é inferir a relevância dos termos, dando maior peso àqueles que aparecem mais raramente no conjunto de textos selecionados.

Na última etapa do módulo é realizada a **busca por correspondências** entre os vetores de características textuais. Utilizou-se a técnica OOS (Seção 2.2.3), em que a comparação é feita entre os segmentos de pares de vídeos e a medida de similaridade é o cosseno. Tal técnica necessita que seja definido um limiar para indicar uma possível correspondência. A análise da base indicou que o valor ótimo para esse limiar é proporcional ao tamanho da entrada. Então, fixou-se um tópico (“Encontro dos presidentes Bolsonaro e Trump”) para um ajuste mais fino do valor e propôs-se o cálculo para automatizar a operação, apresentado na Equação 5.1. Nela, a constante 0,17 corresponde ao limiar capaz de gerar os melhores resultados para o conjunto em questão, sendo 785 segundos seu tempo total. Assim, dada uma nova coleção de vídeos, basta utilizar sua duração na Equação 5.1 para obter o limiar mais adequado na busca por correspondências.

$$limiar = 0,17 + 0,17 \times \frac{tempo_{videos} - 785}{785} \quad (5.1)$$

Considerando que a base é representativa, contendo tópicos de diferentes categorias, sugere-se que a relação entre tamanho do conjunto de vídeos e valor do limiar é apropriada para uso no domínio de vídeos de notícias como um todo.

Por fim, como a aplicação da técnica OOS resulta em pares de conteúdo similar, realiza-se um passo adicional de união dos segmentos para a geração de conjuntos, como ilustrado na Figura 23. Supondo que tenham sido obtidos os seguintes pares de informações redundantes: V1-S2 com V2-S5 e V1-S2 com V3-S4, onde V<núm.>-S<núm.> representam o vídeo e seu respectivo segmento. Nesse passo, realiza-se a união dos dados, gerando o conjunto formado por V1-S2, V2-S5 e V3-S4.

Logo, a saída deste módulo consiste em segmentos considerados redundantes organizados em conjuntos.

⁸ O NLTK é uma plataforma para o desenvolvimento de programas em Python que trabalham com dados em linguagem natural. Fornece interfaces para mais de 50 corpora e recursos lexicais, junto com um pacote de bibliotecas de processamento de texto para classificação, tokenização, *stemming*, marcação, análise e raciocínio semântico (<https://www.nltk.org>).

Figura 23 – Exemplo de união de pares de segmentos redundantes identificados pela técnica OOS. Na imagem denota-se V<núm.> - S<núm.>, representando o vídeo e seu respectivo segmento.

Fonte: Elaborada pelo autor.

5.4 Seleção de segmentos

O módulo anterior (Seção 5.3) tem como resultado conjuntos de segmentos redundantes. Logo, é necessário escolher um representante de cada conjunto para ser levado ao sumário, de maneira a mantê-lo compacto e sem repetição de informações.

Figura 24 – Visão geral do módulo para seleção de segmentos.

Fonte: Elaborada pelo autor.

Considera-se que o conteúdo textual dos segmentos pertencentes a um mesmo conjunto é equivalente, uma vez que esse foi o critério adotado para a identificação de redundância. Assim, a seleção entre as alternativas para compor o sumário é realizada com base no conteúdo visual. Essa solução contribui para a geração de resultados mais informativos, uma vez que são escolhidos segmentos cujos quadros contêm elementos relevantes para um melhor entendimento da notícia, complementando o que é relatado no texto. Para isso, foi desenvolvido um modelo de representação capaz de evidenciar as características visuais mais representativas e permitir comparação. Na Figura 24, ilustram-se as três etapas principais deste módulo: 1) extração de quadros-chave; 2) *Bag of Features*; 3) escolha dos segmentos, as quais são discutidas a seguir.

Extração de quadros-chave

São obtidos quadros-chave dos segmentos classificados como redundantes. Para isso, empregou-se o método KS-SIFT (Seção 2.1.3), cujo princípio é verificar se uma imagem candidata é suficientemente diferente de todas as já selecionadas. Então, são extraídos dos

quadros-chave vetores de características SIFT (Seção 2.2.2), resultando em um conjunto de descritores para cada segmento. Foram utilizados recursos da biblioteca OpenCV para leitura das imagens e extração de características.

Bag of Features

Para a representação de cada segmento, utiliza-se o modelo *Bag of Features*, também chamado *Bag of Visual Words* (HERVE; BOUJEMAA, 2009), uma adaptação do *Bag of Words* (Seção 2.2.1) para o contexto de imagens. O primeiro passo é a geração de um dicionário de palavras visuais⁹ por meio de agrupamento, onde o centróide de cada grupo é tomado como termo. Nesse caso, adotou-se o KMeans como método de agrupamento por ser simples e eficaz para o cenário de aplicação, usualmente utilizado nessa etapa do modelo. Como é comum na área de pesquisa, o tamanho do dicionário foi definido empiricamente na base de vídeos, variando-se os valores de 50 a 700, passo 50, e verificando-se o impacto nos resultados. Considerou-se que 300 grupos apresentavam a melhor relação desempenho/eficácia. Utilizou-se a implementação do KMeans disponível na biblioteca *sklearn.cluster*.

O próximo passo é a obtenção de um histograma de palavras visuais para cada segmento. Nesse ponto, diferentemente da abordagem usual em trabalhos relacionados, na qual realiza-se a contagem do número de ocorrências das palavras, propõe-se o cálculo do valor TF-IDF de maneira análoga ao empregado no modelo *Bag of Words*, porém utilizando palavras visuais ao invés de textuais. A vantagem está em incluir uma medida de relevância para os termos, o que será especialmente importante na próxima etapa do módulo (Escolha dos segmentos). Na Equação 5.2, relembra-se a definição do valor TF-IDF, adaptando-a ao contexto de uso. Dado um histograma que representa um segmento de vídeo, o valor de cada posição é obtido pela frequência de tal palavra visual no segmento, multiplicada por uma função inversa do número de segmentos em que ela ocorre.

$$IDF(w) = \log 10 \frac{N}{DF(w)}$$

N : número total de segmentos

$DF(w)$: número de segmentos que contêm a palavra visual w

$TF(w, s)$: número de ocorrências da palavra visual w no segmento s

$TFIDF(w, s) = TF(w, s) \times IDF(w)$

(5.2)

Escolha dos segmentos

Nesta etapa, seleciona-se o segmento mais relevante de cada conjunto de informações redundantes para ser levado ao sumário. A relevância é avaliada pelo somatório dos valores TF-IDF dos histogramas de palavras visuais. Pela própria definição do TF-IDF, ao adotar tal abordagem, se está ponderando a frequência das palavras visuais por quão raras elas são

⁹ Palavras visuais são definidas como características representativas do conteúdo da base de imagens (HERVE; BOUJEMAA, 2009).

na coleção de vídeos, logo, terão mais importância segmentos com mais conteúdo visual, porém de maior relevância. Para cada conjunto de informações redundantes, o segmento que possuir histograma com maior valor do somatório será selecionado para compor o sumário.

Portanto, neste módulo, incorporou-se o critério de exclusão redundância, visando evitar a inclusão de informações repetidas no sumário. A escolha de uma opção de cada conjunto de segmentos redundantes, por sua vez, pautou-se no critério conteúdo visual.

5.5 Produção do sumário

Este módulo aborda a inclusão no sumário dos segmentos selecionados até aqui considerando os critérios cronologia e fluência. Quanto à cronologia, o maior desafio em se trabalhar com multivídeo é organizar subtópicos provenientes de diversos vídeos, já que um mesmo assunto pode ser abordado em posições diferentes dependendo de sua fonte.

Para lidar com esse problema, propõe-se utilizar informações dos conjuntos de segmentos redundantes. Conforme exemplificado na Figura 25, supõe-se que o sumário já possui o item V2-S3 e deseja-se inserir V1-S4, onde V<núm.>-S<núm.> denota o vídeo e seu respectivo segmento. É possível observar que se tratam de segmentos provenientes de dois vídeos-fonte. Considera-se que o item V1-S4 pertence ao conjunto de segmentos redundantes formado por: V1-S4 e V2-S2, indicando que discorrem sobre o mesmo assunto. Esse dado permite que, para determinar a posição de inserção, a comparação seja feita entre subtópicos do mesmo vídeo, isto é, pode-se comparar V2-S3 com V2-S2. Logo, o item V1-S4 será adicionado ao sumário antes de V2-S3.

Figura 25 – Exemplo de inserção de segmentos no sumário respeitando a ordem cronológica. Na imagem denota-se V<núm.> - S<núm.>, representando o vídeo e seu respectivo segmento.

Fonte: Elaborada pelo autor.

Quando o conjunto de informações redundantes não possui um item do mesmo vídeo do item já incluído no sumário, as posições dos segmentos (a inserir e do sumário) são comparadas diretamente. Porém, essa solução para os casos em que faltam dados para propiciar uma análise mais detalhada dos subtópicos pode ocasionar erros de ordenação uma vez que, como já mencionado, um assunto não é abordado necessariamente na mesma posição em fontes diferentes.

O vídeo do sumário é produzido pela concatenação da introdução e dos demais subtópicos, selecionados nas etapas anteriores, inseridos de acordo com a ordem cronológica.

Para lidar com a fluência, propõe-se adicionar efeitos de *fade-in* e *fade-out* nas transições de vídeo e áudio entre os segmentos. Trata-se de um curto período de tempo (neste trabalho, utilizou-se 0,5 segundo) em que a imagem de um segmento fica gradualmente mais escura e a do próximo é apresentada nesse tom mais escuro sendo, aos poucos, revelada. Similarmente, a intensidade do som é reduzida e o áudio do próximo segmento é gradativamente introduzido. Tais efeitos contribuem para uma melhor experiência de visualização dos usuários, uma vez que os cortes não são abruptos e se favorece a percepção de continuidade. Utilizou-se a biblioteca *moviepy.editor* para a geração do resultado (recorte dos vídeos originais, concatenação de segmentos, adição de efeitos para fluência e gravação do arquivo de vídeo contendo o sumário).

5.6 Considerações finais

Neste capítulo, apresentou-se a técnica de sumarização multivídeo HS-MVideoSumm, que se baseia em estratégias humanas. Na Tabela 10, é possível observar um resumo dos critérios identificados no estudo com usuário, discutido no Capítulo 4, que foram incorporados na técnica.

A seguir, indicam-se os módulos da técnica HS-MVideoSumm em que cada critério destacado na Tabela 10 foi abordado.

- Introdução: módulo “Seleção da introdução” (Seção 5.1).
- Cronologia e fluência: módulo “Produção do sumário” (Seção 5.5).
- Redundância - critério de inclusão: módulo “Detecção de segmentos redundantes” (Seção 5.3).
- Cobertura: a preocupação em manter o sumário o mais completo possível foi considerada de modo geral pela técnica, inserindo-se todos os segmentos identificados como relevantes (por exemplo, “Seleção da introdução” - Seção 5.1, “Detecção de segmentos redundantes” - Seção 5.3, “Seleção de segmentos” - Seção 5.4). Somente foram utilizados filtros baseados em critérios provenientes do usuário, não foram incorporadas restrições adicionais que poderiam exigir o corte arbitrário de informações, por exemplo, o tempo máximo do sumário.

Tabela 10 – Estratégias humanas para sumarização multivídeo identificadas no estudo com usuários. Os critérios destacados em negrito foram incorporados na técnica HS-MVideoSumm.

Critérios de inclusão
Introdução
Cronologia
Fluência
Redundância
Informação única
Cobertura
Conteúdo visual
Segmentos com fonte de informação
Critérios de exclusão
Subjetividade
Informações complementares
Redundância
Comentários ou opiniões de jornalistas
Entrevistas

Fonte: Dados da pesquisa.

- Conteúdo visual: módulo “Seleção de segmentos” (Seção 5.4). Esse foi o critério que guiou a escolha de um entre os segmentos redundantes.
- Subjetividade, comentários ou opiniões de jornalistas e entrevistas: módulo “Detecção de subjetividade” (Seção 5.2).
- Redundância - critério de exclusão: módulo “Seleção de segmentos” (Seção 5.4).

A técnica HS-MVideoSumm incorporou a maioria dos critérios para seleção de conteúdo revelados no estudo com usuários. Além disso, foi desenvolvida de maneira modular, o que facilita a inclusão de novos critérios, bem como alterações nas soluções propostas.

AVALIAÇÃO

Neste capítulo, apresenta-se a avaliação da técnica HS-MVideoSumm e respectivos módulos por meio de experimentos objetivos e subjetivos. Na Seção 6.1, introduz-se a base de vídeos elaborada no domínio de notícias. Na Seção 6.2, são discutidos experimentos realizados visando verificar a eficácia dos módulos da técnica HS-MVideoSumm. Para isso, a avaliação baseia-se no uso de *ground truths* e cálculo de medidas adotadas na literatura da área (precisão, abrangência e F1). Na Seção 6.3, apresentam-se experimentos desenvolvidos com o objetivo de obter a visão dos usuários finais sobre os resultados da sumarização multivídeo. Nesse contexto, destaca-se a comparação dos sumários gerados pela HS-MVideoSumm com trabalhos relacionados, utilizando-se uma abordagem baseada em notas para capturar as preferências dos participantes. Por fim, na Seção 6.4, são pontuadas considerações sobre os experimentos de avaliação.

6.1 Base de vídeos

Bases de vídeo bem estabelecidas na literatura ainda são um problema em aberto na área de sumarização multivídeo. O desafio aumenta ao se acrescentarem restrições como domínio (notícias) e idioma (português). Particularmente o idioma foi uma especificidade necessária no contexto deste trabalho, pois a extração de critérios para seleção de conteúdo dependia da observação dos usuários, o que, com os recursos disponíveis, era viável por meio de encontros presenciais. Por isso, desenvolveu-se uma base para ser utilizada na etapa de avaliação. Vale ressaltar que, como parte dos experimentos é subjetivo, considerou-se conveniente utilizar notícias mais recentes do que as contidas na base descrita na Tabela 6, visando tornar o contexto mais interessante aos participantes. Além disso, como as estratégias humanas foram derivadas do conteúdo da Tabela 6, julgou-se apropriado, nesta etapa, utilizar vídeos diferentes, de maneira a confirmar a validade dos critérios para o domínio.

Tabela 11 – Descrição da base de vídeos utilizada nos experimentos de avaliação.

Telejornal	Título da notícia	Data de exibição	Tempo vídeo*	Link
Apresentação do bebê da família real britânica				
Jornal Nacional	Príncipe Harry e duquesa Meghan Markle apresentam filho recém-nascido	08/05/2019	1:09	Globo Play
Jornal da Record	Novo integrante da Família Real britânica ganha nome de Archie Harrisona	08/05/2019	1:24	Site Record
Jornal da Band	Filho de Harry e Meghan se chama Archie Harrison	08/05/2019	0:51	Site Band
Encontro dos presidentes Bolsonaro e Trump				
Jornal Nacional	Brasil abre mão de tratamento especial na OMC para fazer parte da OCDE	19/03/2019	3:52	Globo Play
Jornal da Record	Jair Bolsonaro se encontra com Donald Trump em Washington	19/03/2019	3:00	Site Record
Jornal da Band	Bolsonaro é recebido por Trump na Casa Branca	19/03/2019	6:13	Site Band
Rompimento da barragem de rejeitos em Brumadinho				
Jornal Nacional	Rompimento de barragem de rejeitos de minério provoca desastre em Brumadinho (MG)	25/01/2019	5:27	Globo Play
Jornal da Record	Bombeiros seguem nas buscas por sobreviventes em Brumadinho (MG)	25/01/2019	7:32	Site Record
Jornal da Band	Bombeiros falam em até 300 desaparecidos em Brumadinho	25/01/2019	4:03	Site Band
Incêndio na Catedral de Notre-Dame				
Jornal Nacional	Incêndio consome catedral de Notre-Dame de Paris	15/04/2019	2:42	Globo Play
Jornal da Record	Catedral de Notre-Dame é destruída por incêndio em Paris	15/04/2019	2:32	Site Record
Jornal da Band	Paris: Incêndio consome Catedral de Notre-Dame	15/04/2019	3:14	Site Band
Relação comercial entre Google e Huawei				
Jornal Nacional	Google rompe com Huawei e celulares não terão atualização do Android	20/05/2019	1:43	Globo Play
Jornal da Band	Google não vai atualizar aplicativos dos celulares Huawei	20/05/2019	0:53	Site Band
SBT Brasil	Google suspende sistema Android em celulares de marca chinesa	20/05/2019	1:43	YouTube

*Tempo em minutos.

Fonte: Dados da pesquisa.

Foram selecionados cinco tópicos de destaque na mídia internacional, pertencentes a diferentes categorias (política, tragédia, curiosidade e tecnologia) e com durações variadas. Cada um deles é composto por três vídeos originalmente transmitidos em telejornais brasileiros bem

estabelecidos e conceituados. Na Tabela 11, apresentam-se as principais características da base. Por serem provenientes de várias fontes, os vídeos foram padronizados quanto a: formato (mp4), taxa de quadros (25 *fps*) e resolução (1280x720 *pixels*), por meio do *framework* FFmpeg.

A técnica HS-MVideoSumm processa, além do conteúdo visual, informações textuais. Como os vídeos selecionados não possuíam legendas, aplicou-se a *Web Speech API*¹, uma ferramenta para Reconhecimento Automático de Fala (do inglês, *Automatic Speech Recognition* (ASR)), para realizar a transcrição do áudio. Para que erros no processo de transcrição não influenciassem negativamente os resultados, os textos obtidos passaram por uma etapa de correção manual. Nela, foram substituídas palavras incorretamente identificadas e acrescentou-se pontuação.

Em seguida, para a tarefa de segmentação em subtópicos, unidade básica de processamento da técnica HS-MVideoSumm, foram recrutados 36 usuários. O convite foi realizado pessoalmente ou por meios eletrônicos e o perfil definido restringiu pessoas com ensino médio como escolaridade mínima, de maneira que a compreensão do conteúdo não impusesse dificuldades. Cada participante recebeu um texto de notícia² e o respectivo *link* do vídeo de onde ele foi obtido. Explicou-se que a tarefa era dividir esse texto em subtópicos. O objetivo era agrupar as frases relacionadas, que abordavam um mesmo assunto, contido no tópico principal. Cada texto foi segmentado por três usuários e obteve-se um consenso pela análise das transições, isto é, se duas ou mais pessoas indicaram uma mudança de subtópico após determinada frase, essa transição foi incluída no consenso.

É importante destacar que os segmentos textuais identificados pelos usuários possuem conteúdos visuais associados. Para obtê-los, foi realizada manualmente a anotação dos tempos de início e fim de cada subtópico em seu respectivo vídeo, gerando-se arquivos JSON. Na Figura 26, apresenta-se um exemplo de organização das informações derivadas do processo de segmentação.

Portanto, compõem a base: 1) conjuntos de vídeos de notícias padronizados; 2) seus respectivos textos segmentados em subtópicos. Na Tabela 12, apresenta-se o número de segmentos identificados para cada tema.

Vale ressaltar que o tamanho da base, bem como o número de vídeos por conjunto foi definido considerando-se que interações humanas seriam necessárias tanto na etapa de preparação dos vídeos, como para avaliação dos resultados. Nesse sentido, foi fundamental garantir a viabilidade em termos de tempo demandado do usuário.

¹ A *Web Speech API* pode ser acessada, pelo navegador Google Chrome, por meio do seguinte endereço: www.google.com/chrome/demos/speech.html.

² Como os tópicos “Apresentação do bebê da família real britânica” e “Relação comercial entre Google e Huawei” possuíam pouco conteúdo, nesses casos, um mesmo usuário segmentou dois textos (um de cada conjunto).

Figura 26 – Exemplo de arquivo JSON com a segmentação de um vídeo em subtópicos. Contém os campos: 1) *content* (texto do segmento); 2) *begin* (tempo de início do segmento); 3) *end* (tempo final do segmento).

```
[
  {
    "content": "O príncipe Harry e a duquesa Meghan Markle apresentaram hoje o filho recém-nascido.",
    "begin": "00:00:00",
    "end": "00:00:06"
  },
  {
    "content": "A primeira aparição de Harry foi aos olhos de todos, uma multidão se acotovelou em frente ao hospital. O primeiro vislumbre do filho do príncipe foi só para um seleto grupo da imprensa e, junto com o novo bebê real, o casal apresentou um estilo discreto. O menino estava embulhado em carinhos.",
    "begin": "00:00:06",
    "end": "00:00:26"
  },
  {
    "content": "Meghan Markle falou da magia que é virar mãe, sobretudo por ter dois dos melhores caras do mundo. O bebê vem fazendo por merecer, temperamento calmo e doce. Harry disse que nem sabe a quem ele puxou.",
    "begin": "00:00:26",
    "end": "00:00:42"
  },
  {
    "content": "O casal também quebrou hoje outro mistério e revelou o nome do novo bebê real: Archie Harrison Mountbatten-Windsor, um nome até bem mais curto do que os padrões da realeza.",
    "begin": "00:00:42",
    "end": "00:00:52"
  },
  {
    "content": "Elizabeth Alexandra Mary conheceu hoje o bisneto. A rainha apareceu ao lado do marido e da mãe de Meghan.",
    "begin": "00:00:52",
    "end": "00:00:59"
  },
  {
    "content": "O casal ainda não sabe com quem Archie se parece. O príncipe disse que o rosto de bebê muda muito, só que o de pai e mãe muda ainda mais.",
    "begin": "00:00:59",
    "end": "00:01:09"
  }
]
```

Fonte: Dados da pesquisa.

Tabela 12 – Tempo total e número de segmentos de cada conjunto de vídeos.

Tópico	Tempo (min)	Segmentos
Apresentação do bebê da família real britânica	03:24	16
Encontro dos presidentes Bolsonaro e Trump	13:05	29
Rompimento da barragem de rejeitos em Brumadinho	17:02	47
Incêndio na Catedral de Notre-Dame	08:28	29
Relação comercial entre Google e Huawei	04:19	16

Fonte: Dados da pesquisa.

6.2 Avaliação objetiva

Para os módulos em que era aplicável (“[1] Seleção da introdução”, “[2] Detecção de subjetividade”, “[3] Detecção de segmentos redundantes”) foram desenvolvidos experimentos com o objetivo de verificar a eficácia das soluções propostas quanto à identificação das informa-

ções de interesse. A base para comparação foram *ground truths* manualmente desenvolvidos, para cada módulo avaliado, os quais são detalhados nas Seções 6.2.1 a 6.2.3. Vale mencionar que as informações classificadas nos vídeos (introdução, comentários ou opiniões de jornalistas, entrevistas e redundância) não tinham natureza ambígua, logo, os *ground truths* elaborados não possuem viés que justifique a produção de diversas versões e posterior consenso.

6.2.1 Detecção da introdução

Para a realização deste experimento, isolou-se o módulo “[1] Seleção da introdução”, recebendo como entrada os vídeos e segmentações em subtópicos. Então, analisou-se se a solução proposta foi capaz de identificar corretamente: 1) o quadro final da introdução e 2) o número de segmentos que a compõem, partindo-se do início de cada vídeo. Um trecho do *ground truth* desenvolvido é apresentado na Tabela 13, onde se pode observar o último quadro pertencente à introdução, bem como o respectivo número de segmentos, para cada vídeo relacionado ao incêndio na Catedral de Notre-Dame.

Tabela 13 – *Ground truth* desenvolvido para avaliação do módulo “[1] Seleção da introdução”. O trecho apresentado contém as informações para o tópico “Incêndio na Catedral de Notre-Dame”.

	Quadro final	Núm. segmentos
Jornal Nacional	18	1
Jornal da Band	21	1
Jornal da Record	30	3

Fonte: Dados da pesquisa.

Com o objetivo de verificar a eficácia do módulo, foram calculadas as porcentagens de acertos, considerando-se todos os tópicos da base (Tabela 11), para as duas configurações de avaliação: **quadro final da introdução** - 80% e **segmentos que compõem a introdução** - 93%.

Os resultados mostram que a abordagem proposta atinge altas taxas de acerto, principalmente para o nível de segmento, unidade processada pela técnica HS-MVideoSumm. Além disso, confirmam o padrão visual dos quadros de introdução no domínio de notícias, sendo o uso de histogramas um recurso eficaz e eficiente para este caso.

Uma limitação do módulo envolve vídeos que não possuem introdução. Tal situação não é tratada explicitamente, mas é parte da solução devido às características consideradas. Quando não há a apresentação da notícia pelo âncora, a tendência é que exista desde o início uma maior variação no conteúdo visual dos quadros, levando o algoritmo a identificar um quadro final da introdução. Entretanto, de modo geral, esse quadro final não irá corresponder ou ultrapassar o limite do segmento, evitando que seja classificado como introdução. Ou seja, apesar da identificação incorreta do quadro de transição, a abordagem é aplicável a nível de segmento, não prejudicando a geração do sumário.

6.2.2 Detecção de subjetividade

Este experimento foi realizado com o objetivo de verificar a eficácia do módulo quanto à detecção de subjetividade. Para isso, foram utilizados *ground truths* para comparação dos resultados produzidos pela abordagem proposta. Na Figura 27, um trecho do *ground truth* elaborado para o vídeo do telejornal SBT Brasil, tópico “Relação comercial entre Google e Huawei”, é apresentado. Pode-se observar que foi acrescentado um campo (*class*) ao arquivo JSON que contém a segmentação em subtópicos. Tal campo foi manualmente anotado com o valor “neutro” ou “não neutro”, no último caso indicando uma opinião de jornalista ou entrevista.

Figura 27 – *Ground truth* desenvolvido para avaliação do módulo “[2] Detecção de subjetividade”. O trecho apresentado contém a classificação de alguns segmentos do vídeo do telejornal SBT Brasil, tópico “Relação comercial entre Google e Huawei”.

```
{
  "content": "O Google suspendeu o uso do sistema Android em celulares da marca Huawei. A decisão segue uma ordem do presidente dos Estados Unidos.",
  "begin": "00:00:00",
  "end": "00:00:10",
  "class": "neutro"
},
{
  "content": "A briga é entre os Estados Unidos e a China, mas a repercussão mundial. \"Acho um absurdo, porque o telefone funciona bem e é eficiente\", disse o italiano que usa um smartphone da marca chinesa.",
  "begin": "00:00:10",
  "end": "00:00:23",
  "class": "não neutro"
},
{
  "content": "Na semana passada, Donald Trump assinou um decreto alegando emergência nacional para impedir que empresas americanas negociem com algumas companhias chinesas, entre elas a Huawei. O governo dos Estados Unidos afirma que a China está usando os equipamentos da segunda maior fabricante de smartphones do mundo para espionagem.",
  "begin": "00:00:23",
  "end": "00:00:45",
  "class": "neutro"
},
}
```

Fonte: Dados da pesquisa.

Foram calculadas medidas de eficácia usualmente empregadas na área de pesquisa, isto é, precisão, abrangência e F1 (Seção 2.4.1) para: 1) todos os segmentos de cada conjunto de vídeos; 2) etapa atual da técnica HS-MVideoSumm, ou seja, entrada do módulo considerando os processamentos anteriormente realizados. Para o caso de “[2] Detecção de subjetividade”, as introduções não são consideradas por já terem sido selecionadas/excluídas, logo, as medidas são computadas apenas sobre os segmentos remanescentes. Isso é feito para possibilitar tanto uma análise abrangente da solução quanto aplicada ao contexto de uso. Na Tabela 14, apresentam-se os resultados. Os valores indicam a média de cada métrica para os três vídeos de cada conjunto.

Analisando-se os resultados, pode-se verificar que, quando todos os segmentos são considerados, a abordagem alcança, de modo geral, altos valores de precisão e abrangência. Na base de vídeos completa, a média da medida F1 é 77%, indicando que é possível identificar a maioria dos segmentos contendo subjetividade, com poucos deles incorretamente classificados. Ao avaliar apenas as informações de entrada do módulo, a média da medida F1 para toda a

Tabela 14 – Avaliação da técnica HS-MVideoSumm quanto à detecção de subjetividade. São apresentadas as métricas precisão (P), abrangência (A) e F1 calculadas sobre: 1) todos os segmentos de cada conjunto e 2) entrada do módulo “[2] Detecção de subjetividade” da técnica HS-MVideoSumm.

Tópico	Todos os segmentos			Módulo HS-MVideoSumm		
	P	A	F1	P	A	F1
Apresentação do bebê da família real britânica	1,00	0,89	0,93	1,00	0,89	0,93
Encontro dos presidentes Bolsonaro e Trump	0,93	0,78	0,83	0,93	0,78	0,83
Rompimento da barragem de rejeitos em Brumadinho	0,54	1,00	0,69	0,90	1,00	0,94
Incêndio na Catedral de Notre-Dame	0,72	0,78	0,73	1,00	0,78	0,83
Relação Comercial entre Google e Huawei	0,67	0,67	0,67	1,00	0,67	0,67
	0,77	0,82	0,77	0,97	0,82	0,84

Fonte: Dados da pesquisa.

base sobe para 84%. Pode-se observar, na Tabela 14, que essa diferença se deve a melhorias na precisão, evidenciando que alguns segmentos de introdução haviam sido rotulados como conteúdo subjetivo. Esse é um resultado esperado uma vez que, em especial para assuntos de maior repercussão, os jornalistas demonstram seus sentimentos a respeito do acontecimento que será noticiado, por exemplo, indicando choque, reprovação, tristeza, entre outros. Além disso, a melhoria na medida de avaliação ressalta a importância de realizar os processamentos de forma modular, possibilitando que a etapa atual se concentre nos segmentos de interesse, sem a interferência de informações não necessárias ao contexto.

Vale mencionar que a detecção de subjetividade é influenciada por características da base de vídeos. Utilizam-se como entrada subtópicos identificados pelos usuários. Portanto, se uma opinião ou entrevista representar uma pequena parcela do segmento, o conteúdo pode não ser considerado subjetivo o suficiente e ser rotulado como neutro. Porém, pelos resultados pode-se perceber que essa limitação se refere a casos de exceção e a solução proposta é apropriada para ser aplicada no domínio de vídeos de notícias, apresentando alta acurácia e cobertura.

6.2.3 Detecção de segmentos redundantes

De modo análogo às Seções 6.2.1 e 6.2.2, o objetivo deste experimento é medir a eficácia do módulo. Neste caso, é relevante considerar que, para cada vídeo de um tópico, podem existir diferenças em como os assuntos foram distribuídos em segmentos pelos usuários, o que gera diversas possibilidades de agrupamento de informações redundantes. Por isso, como *ground truth*, para cada tópico da base, elaborou-se uma lista de assuntos abordados em mais de um vídeo. Assim, é possível avaliar se o módulo é capaz de gerar conjuntos de segmentos que

incluam os itens dessa lista. Na Tabela 15, apresenta-se o *ground truth* desenvolvido para o tópico “Relação comercial entre Google e Huawei”.

Tabela 15 – *Ground truth* desenvolvido para avaliação do módulo “[3] Detecção de segmentos redundantes”. O trecho apresentado contém os assuntos abordados em mais de um vídeo para o tópico “Relação comercial entre Google e Huawei”.

Assuntos redundantes
Controle sobre a rede 5G
Estados Unidos acusam a Huawei de espionagem
Celulares da Huawei não terão mais aplicativos da Google como YouTube
Ordem do presidente dos Estados Unidos proibindo negócios com a Huawei
Huawei é a segunda maior fabricante de <i>smartphones</i> do mundo

Fonte: Dados da pesquisa.

Foram elaborados *ground truths* conforme exemplificado na Tabela 15 e calculadas medidas de precisão, abrangência e F1 para: 1) todos os segmentos de cada conjunto de vídeos; 2) etapa atual da técnica HS-MVideoSumm. Para a segunda configuração, considera-se como entrada do módulo “[3] Detecção de segmentos redundantes” o caso ideal, isto é, introduções e opiniões de jornalistas/entrevistas são excluídas com base em seus respectivos *ground truths*. Na Tabela 16, apresenta-se a média de cada medida de avaliação para os três vídeos de cada conjunto, considerando-se todos os seus segmentos.

Tabela 16 – Avaliação da técnica HS-MVideoSumm quanto à detecção de redundância considerando-se todos os segmentos de cada conjunto de vídeos.

Tópico	P	A	F1
Apresentação do bebê da família real britânica	1,00	1,00	1,00
Encontro dos presidentes Bolsonaro e Trump	0,88	0,78	0,82
Rompimento da barragem de rejeitos em Brumadinho	0,80	0,67	0,73
Incêndio na Catedral de Notre-Dame	0,89	1,00	0,94
Relação Comercial entre Google e Huawei	1,00	1,00	1,00
	0,91	0,89	0,90

Fonte: Dados da pesquisa.

Analisando-se a Tabela 16, é possível verificar que, para o caso em que todos os segmentos são utilizados, a abordagem proposta atingiu 90% de média para a medida F1, indicando que possui alta eficácia para detecção de redundância no domínio.

Considerando-se apenas os dados de entrada do módulo, são gerados conjuntos que incorporam todos os assuntos redundantes, sem grupos adicionais irrelevantes, logo, a média da medida F1 para toda a base de vídeos é 100%. Isso evidencia que filtrar informações em passos anteriores é importante para eliminar ruídos, restringir o contexto de aplicação do módulo e aprimorar os resultados.

6.3 Avaliação com usuários

Esta seção apresenta uma análise da técnica HS-MVideoSumm sob dois diferentes aspectos. No primeiro, busca-se avaliar os resultados da sumarização ao se aplicar (ou não) o critério de exclusão “subjetividade” (Seção 6.3.1). Vale ressaltar que se optou por aprofundar o estudo deste módulo em particular por ser uma tarefa mais desafiadora. No segundo, faz-se uma comparação dos sumários produzidos pela HS-MVideoSumm a trabalhos relacionados (Seção 6.3.2). Nos dois casos, o método selecionado para realizar a análise é a Avaliação Subjetiva, uma vez que se espera obter a opinião do usuário e verificar se os resultados atendem às suas expectativas.

6.3.1 Detecção de subjetividade

O objetivo desta avaliação é analisar o impacto que a remoção de segmentos com conteúdo subjetivo tem nos sumários multivídeo e qual a percepção dos usuários. Pretende-se confirmar se o filtro derivado de estratégias humanas é capaz de gerar resultados alinhados aos julgamentos de relevância dos usuários finais.

Utilizando-se a técnica de sumarização multivídeo HS-MVideoSumm (Figura 16) e a base de vídeos de notícias (Tabela 11) foram gerados sumários com duas configurações: 1) removendo-se os segmentos subjetivos dos resultados; 2) mantendo-se o conteúdo subjetivo nos sumários finais. Em outras palavras, no primeiro caso, executou-se a técnica completa, com todas as etapas descritas na Figura 16. Já no segundo caso, o módulo “[2] Detecção de subjetividade” foi suprimido, sendo que o módulo seguinte (“[3] Detecção de segmentos redundantes”) recebeu a saída do anterior (“[1] Seleção da introdução”), isto é, os segmentos após a remoção apenas das introduções.

Vale mencionar que, para o tópico “Relação comercial entre Google e Huawei”, como o conjunto de vídeos possui poucos segmentos com opiniões de jornalistas e entrevistas, a remoção não influenciou os resultados da sumarização, desse modo, as duas configurações do experimento produziram sumários idênticos. Por isso, esse tópico não foi incluído nas discussões a seguir.

A avaliação subjetiva foi baseada na abordagem Matriz de Preferências (Seção 2.4.2), adotada por trabalhos relacionados (IDE *et al.*, 2017; NIE *et al.*, 2016; ZHANG *et al.*, 2015a). Para a condução do experimento, preparou-se um formulário *on-line* (Apêndice B.1) a ser preenchido pelos usuários. Nele, primeiramente foi fornecida uma breve descrição da atividade. Então, para cada tópico de notícias, foram apresentadas duas opções de sumários de vídeo e solicitou-se que os participantes expressassem suas preferências pela escolha de um resultado ou indicando que não consideravam as diferenças relevantes. A ordem de apresentação dos sumários com e sem segmentos subjetivos foi previamente definida de maneira aleatória. No total, exibiu-se em torno de 16 minutos de vídeo aos usuários e o tempo estimado de cada sessão de avaliação foi de 20 a 25 minutos.

Além disso, os participantes não foram informados sobre o processo de sumarização, ou seja, de que um sumário continha segmentos subjetivos e outro não. Assim, após a realização das tarefas, por meio do mesmo formulário, perguntou-se (em uma caixa de texto) quais critérios eles haviam considerado para fazer as escolhas. Desse modo foi possível analisar se as pessoas perceberam o aspecto que estava sendo investigado e o consideraram importante na tomada de decisão.

Para validar o planejamento, conduziu-se um teste piloto com 3 participantes. Solicitou-se que preenchessem o formulário de avaliação dos sumários e, então, respondessem a algumas perguntas sobre a atividade realizada, descritas a seguir.

- A apresentação do formulário permitiu compreender com clareza a tarefa a ser realizada?

Resposta: sim (3 usuários).

- Foi possível entender o que são sumários multivídeo e como foram gerados?

Resposta: sim (3 usuários).

- As informações de contexto fornecidas foram suficientes para que você escolhesse entre as opções apresentadas? Ou você acha que seria necessário assistir a todos os vídeos originais para tomar uma melhor decisão?

Resposta: sim (3 usuários).

- Qual foi o critério que você utilizou para fazer a escolha entre as opções apresentadas?

Respostas: sumário mais conciso, mais direto, com menos repetição, mais organizado, com mais informação, menos sentimento e menos pessoas falando.

- Sem a alternativa “Não considero as diferenças entre os vídeos relevantes”, você seria capaz de optar por um dos dois sumários em todos os temas?

Resposta: não (3 usuários).

Essa entrevista pós-sessão permitiu confirmar que a descrição da tarefa estava elucidativa e que as informações de contexto eram suficientes, não suscitando dúvidas. Também concluiu-se que as pessoas foram capazes de entender as diferenças entre os sumários de vídeo apresentados ao realizar as escolhas.

Em seguida, desenvolveu-se um estudo com 15 usuários. O recrutamento foi feito por meio de um convite *on-line*. Os participantes possuíam o seguinte perfil: ensino médio como escolaridade mínima, não especialistas em sumarização de vídeo, idades entre 20 e 55 anos, 8 homens e 7 mulheres.

Na Figura 28 apresentam-se os resultados da avaliação. Para cada tópico, os participantes puderam escolher entre três opções: 1) sumário com segmentos subjetivos; 2) sumário sem seg-

mentos subjetivos; 3) diferenças irrelevantes entre os sumários. Como explicado anteriormente, eles não estavam cientes das configurações utilizadas no processo de sumarização.

Figura 28 – Preferências dos usuários quanto aos sumários multivídeo.

Fonte: Dados da pesquisa.

É importante destacar que o impacto que a remoção de segmentos subjetivos tem nos resultados pode variar dependendo das características do conjunto de vídeos. Quando um tópico de entrada possui várias opiniões de jornalistas e entrevistas, a exclusão desse conteúdo se reflete mais claramente nos sumários. Por outro lado, quando existem poucos segmentos subjetivos, sua remoção pode não ser evidente nos resultados da sumarização.

Considerando essa situação, pode-se dividir a análise das preferências dos usuários em dois grupos. Os tópicos “Apresentação do bebê da família real britânica” e “Encontro dos presidentes Bolsonaro e Trump” são casos em que os sumários (com e sem segmentos subjetivos) possuem diferenças evidentes. Para eles, como mostrado na Figura 28, a maioria dos usuários (87% e 66%, respectivamente) prefere os resultados sem opiniões de jornalistas e entrevistas, o que é confirmado pelas justificativas das escolhas. Os participantes apontaram que os critérios que utilizaram para decidir entre as opções foram: “menos repetição de informação”, “melhor organização”, “entendimento da notícia”, “apresentação das informações mais importantes”, “vídeo mais compacto”, “sumário sem conteúdo irrelevante ou emocional”, “fatos reportados com mais objetividade”. Tais respostas estão relacionadas com o aspecto em investigação, indicando que as pessoas perceberam diferenças no conteúdo dos sumários de vídeo apresentados e

demonstraram preferência pelos mais objetivos.

O segundo grupo é composto pelos tópicos “Rompimento da barragem de rejeitos em Brumadinho” e “Incêndio na Catedral de Notre-Dame”, os quais contêm poucas opiniões de jornalistas e entrevistas. Para esses casos, os sumários gerados apresentam variações sutis (com e sem segmentos subjetivos) e a maioria dos usuários indicou não considerar as diferenças relevantes (67% e 47%, respectivamente). As demais decisões feitas pelos participantes podem ser justificadas por outros aspectos dos sumários, como a ordenação das informações.

Portanto, conclui-se que opiniões de jornalistas e entrevistas são fortes candidatas a serem eliminadas dos sumários, contribuindo para mantê-los compactos e informativos. As preferências dos participantes indicaram que as informações removidas não eram essenciais. Pelo contrário, em alguns casos, tornavam os sumários menos interessantes na visão dos usuários, com repetição de conteúdo ou falta de objetividade.

6.3.2 Comparação entre HS-MVideoSumm e trabalhos relacionados

Apesar de algumas técnicas de sumarização multivídeo terem sido desenvolvidas nos últimos anos, conforme discutido no Capítulo 3, elas não se encontram publicamente disponíveis, o que dificulta a comparação de novas soluções. Além disso, muitas possuem especificidade de domínio, não sendo aplicáveis a vídeos de notícias. Considerando esses fatores, dois trabalhos da literatura puderam ser selecionados para comparação da técnica HS-MVideoSumm: Video-MMR (LI; MERALDO, 2016; LI; MERALDO, 2010b) e RSumm (RIBALDO; CARDOSO; PARDO, 2016). A seguir, justificam-se as escolhas realizadas e discutem-se as configurações utilizadas.

Video-MMR

Propõe uma solução baseada em redundância e diversidade, critérios aplicados pela maioria dos trabalhos relacionados a sumarização multivídeo, representando-os nos experimentos de avaliação. Além disso, é amplamente utilizado para comparação na literatura (WU; ZHONG; LIU, 2020; JI *et al.*, 2020; JI *et al.*, 2019; PANDA; MITHUN; ROY-CHOWDHURY, 2017; ZHANG; ZIMMERMANN, 2016; ZHANG; ZHANG; ZIMMERMANN, 2015; OU *et al.*, 2014) por não possuir limitação de domínio e, apesar de não estar publicamente disponível, apresentar uma abordagem viável de ser reproduzida.

Na Figura 29 sintetizam-se as principais etapas do algoritmo, bem como seus parâmetros. Os quadros extraídos do conjunto de vídeos são representados adotando-se o modelo *Bag of Features* com o descritor SIFT. Em seguida, são calculadas similaridades por meio da medida cosseno. Com tais informações, elege-se o quadro mais parecido com todos os demais como o primeiro a compor o resultado. Por fim, é feita a seleção de quadros semelhantes ao conjunto, mas diferentes dos já incluídos no sumário.

Os autores afirmam que os quadros de vídeo que compõem o sumário podem ser estendidos no tempo para conteúdo dinâmico com áudio. Seguindo esse raciocínio, na avaliação subje-

Figura 29 – Visão geral do método Video-MMR e parâmetros utilizados na implementação.

Fonte: Elaborada pelo autor.

tiva, realizaram a expansão para 1 e 2 segundos, destacando que os usuários preferiram os segmentos mais longos por proporcionarem mais conforto em relação ao áudio. Considerando tal conclusão e o volume de dados no contexto deste trabalho, fixou-se uma taxa de amostragem de um quadro a cada 3 segundos, com respectivos segmentos de 3 segundos representando os intervalos. Além disso, de maneira análoga a uma das configurações utilizadas pelos autores em seus experimentos, estabeleceu-se 10 quadros como tamanho dos sumários. A implementação foi desenvolvida em Python 3 e os parâmetros (dimensão do dicionário de palavras visuais, lambda) foram definidos de acordo com os artigos que a guiaram (LI; MERALDO, 2016; LI; MERALDO, 2010b).

RSumm

Este método foi selecionado por ser um sumarizador multidocumento com resultados que se destacam na literatura e possuir contexto de uso similar à técnica HS-MVideoSumm: textos de notícias em português. Então, a partir do sumário textual, pode-se obter uma versão em vídeo que permita a comparação.

Na Figura 30, apresenta-se a adaptação realizada para aplicar o método RSumm ao domínio de vídeos. Foram utilizados como entrada os textos provenientes dos vídeos da base descrita na Tabela 11. Então, executou-se o método³ na versão baseada no Caminho Denso-Segmentado, com taxa de compressão de 70%, conforme indicado em (RIBALDO; CARDOSO; PARDO, 2016). Isso significa que o tamanho do sumário corresponde a 30% do maior texto do conjunto, sendo o tamanho dado em função do número de palavras. O RSumm gera como saída um sumário textual composto pelas frases selecionadas. Por fim, a versão em vídeo é produzida concatenando-se os trechos correspondentes a tais frases.

³ O RSumm está disponível em: <https://sites.icmc.usp.br/taspardo/sucinto/resources.html>

Figura 30 – Visão geral do método RSumm aplicado ao domínio de vídeos.

Fonte: Elaborada pelo autor.

As técnicas HS-MVideoSumm, RSumm e Video-MMR foram executadas para os conjuntos de vídeos da base apresentada na Tabela 11, gerando três diferentes resultados para cada tópico. Na Tabela 17 é possível observar os tempos dos sumários de vídeo produzidos por cada técnica e utilizados para avaliação, com um total de aproximadamente 16 minutos.

Tabela 17 – Tamanho (em minutos) dos conjuntos de vídeos e dos sumários produzidos por cada técnica avaliada.

Tópico	Total	HS-MVideoSumm	RSumm	Video-MMR
Apresentação do bebê da família real britânica	03:24	00:34	00:21	00:30
Encontro dos presidentes Bolsonaro e Trump	13:05	02:07	01:46	00:30
Rompimento da barragem de rejeitos em Brumadinho	17:02	02:04	02:14	00:30
Incêndio na Catedral de Notre-Dame	08:28	01:57	00:40	00:30
Relação Comercial entre Google e Huawei	04:19	01:18	00:27	00:30

Fonte: Dados da pesquisa.

O experimento com usuários para avaliar os sumários foi realizado por meio de um formulário *on-line* (Apêndice B.2). Nele, primeiramente apresentou-se uma breve descrição da tarefa. Utilizou-se a abordagem baseada em notas (Seção 2.4.2), adotada por diversos trabalhos relacionados (JI *et al.*, 2020; JI *et al.*, 2019; JI *et al.*, 2018; KANEHIRA *et al.*, 2018; PANDA; ROY-CHOWDHURY, 2017; ZHANG; ZIMMERMANN, 2016; ZHANG; ZHANG; ZIMMERMANN, 2015) por ser intuitiva para o usuário e eficaz para medir seu grau de satisfação. Para cada tópico da base, solicitou-se que os participantes assistissem aos sumários de vídeo gerados pelas três técnicas e, considerando o entendimento da notícia e a organização das informações, atribuísssem a cada um deles notas entre 1 (ruim) e 10 (bom), numa escala inteira. A ordem de exibição dos resultados produzidos por cada abordagem foi definida aleatoriamente.

A avaliação foi realizada com 33 usuários com o seguinte perfil: alunos de graduação do curso Bacharelado em Sistemas de Informação; idades entre 18 e 28 anos, com distribuição indicada na Figura 31a; 28 homens e 5 mulheres, conforme ilustrado na Figura 31b.

Figura 31 – Dados demográficos dos participantes do experimento de avaliação da técnica HS-MVideoSumm.

Fonte: Dados da pesquisa.

Na Tabela 18, apresentam-se a média e o desvio padrão das notas que os participantes atribuíram aos sumários de cada abordagem para os diferentes conjuntos de vídeos. Pode-se verificar que a técnica HS-MVideoSumm alcança os melhores resultados com notas, em média, 1,13 pontos maiores que o método RSumm e 6,06 que o Video-MMR. Alguns aspectos do processo de sumarização justificam esse fato e são discutidos a seguir.

Tabela 18 – Média e desvio padrão das notas que os usuários atribuíram aos sumários gerados por cada técnica para os tópicos da base de vídeos.

Tópico	Média $\pm$ Desvio padrão		
	HS-MVideoSumm	RSumm	Video-MMR
Apresentação do bebê da família real britânica	8,21 $\pm$ 1,82	6,15 $\pm$ 1,62	1,42 $\pm$ 1,03
Encontro dos presidentes Bolsonaro e Trump	8,70 $\pm$ 1,49	6,94 $\pm$ 1,80	1,48 $\pm$ 0,91
Rompimento da barragem de rejeitos em Brumadinho	7,76 $\pm$ 1,44	6,76 $\pm$ 2,12	1,91 $\pm$ 1,77
Incêndio na Catedral de Notre-Dame	7,94 $\pm$ 1,87	7,64 $\pm$ 1,56	1,97 $\pm$ 1,47
Relação Comercial entre Google e Huawei	6,33 $\pm$ 1,59	5,79 $\pm$ 2,26	1,88 $\pm$ 1,41
	7,79 $\pm$ 1,64	6,66 $\pm$ 1,87	1,73 $\pm$ 1,32

Fonte: Dados da pesquisa.

- Unidade básica: a técnica HS-MVideoSumm trabalha com subtópicos e o método RSumm com frases, enquanto o Video-MMR utiliza segmentos de tempo fixo (no caso, 3 segundos). Frases ou subtópicos são unidades em que a semântica é respeitada e garante-se o entendimento do conteúdo de áudio, o que não acontece ao se realizar cortes arbitrários por intervalos de tempo.

- Apresentação das informações: as técnicas HS-MVideoSumm e RSumm preocupam-se com cronologia e fluência, enquanto o sumário produzido pelo método Video-MMR é gerado pela simples concatenação de conteúdo, na sequência em que é selecionado. Com a HS-MVideoSumm propõe-se uma abordagem para identificar a ordem cronológica dos segmentos a serem adicionados e utilizam-se efeitos de *fade* no áudio e nas imagens para proporcionar melhor fluência. Já o RSumm se baseia no uso de sentenças de transição para lidar com a cronologia e preservar a coerência.

Apesar da técnica HS-MVideoSumm e do método RSumm compartilharem algumas preocupações no processo de sumarização que contribuem para a geração de melhores resultados, as notas indicam que os critérios de seleção de conteúdo aplicados pela HS-MVideoSumm foram capazes de aprimorar os sumários, proporcionando maior satisfação do usuário.

Na Figura 32 indicam-se quantas vezes (em porcentagem) cada participante preferiu o resultado de cada técnica de sumarização. É possível observar que 79% deles demonstraram predileção pelos sumários produzidos pela HS-MVideoSumm.

Figura 32 – Preferência dos usuários: por participante, quantas vezes (em porcentagem) preferiu o resultado de cada técnica de sumarização.

Fonte: Dados da pesquisa.

Na Figura 33 faz-se outra leitura das preferências dos usuários, indicando a média das notas que cada um deles deu aos resultados das diferentes técnicas. Nesse caso, pode-se verificar que 82% atribuíram notas mais altas aos sumários gerados pela HS-MVideoSumm.

Na Figura 32, nos casos em que se pode observar uma clara preferência dos participantes, isto é, excluindo-se as situações dos usuários 19, 29 e 30, os resultados da técnica HS-MVideoSumm foram escolhidos em 87% das vezes. Além disso, nos 4 casos (usuários 2, 17, 18 e 23) em que os sumários gerados pelo método RSumm foram indicados como melhores, a

Figura 33 – Preferência dos usuários: por participante, média das notas que atribuiu aos resultados de cada técnica de sumarização.

Fonte: Dados da pesquisa.

diferença entre as notas atribuídas para RSumm e HS-MVideoSumm foi pequena (em média, 0,85 pontos).

De maneira geral, os resultados desta avaliação subjetiva apontam uma predileção dos usuários por sumários produzidos por técnicas que se baseiam em aspectos semânticos. Em particular, a preferência demonstrada pelos resultados da técnica HS-MVideoSumm indica que os critérios para seleção de conteúdo derivados de estratégias humanas são soluções efetivas para aproximar a qualidade dos sumários às expectativas dos usuários, tornando-os semanticamente mais relevantes.

6.4 Considerações finais

Neste capítulo descreveu-se a base de vídeos desenvolvida para os experimentos de avaliação, composta por cinco tópicos de destaque na mídia internacional, de diferentes categorias e durações.

Foram apresentados os resultados de avaliações objetivas, os quais demonstraram que as abordagens propostas atingem altos valores de precisão e abrangência para detecção da introdução, de subjetividade e de segmentos redundantes, sendo apropriadas para uso no domínio de vídeos de notícias.

Também foram discutidas avaliações subjetivas realizadas para: 1) investigar o impacto que a remoção de segmentos subjetivos tem nos sumários multivídeo; 2) comparar a técnica HS-MVideoSumm com dois trabalhos relacionados em destaque na literatura. No primeiro caso, os resultados mostraram que as versões sem opiniões de jornalistas e entrevistas são mais interessantes na visão dos usuários por transmitirem o foco da informação, evitando repetição e

conteúdo emocional. No segundo experimento, verificou-se que os sumários produzidos pela técnica HS-MVideoSumm receberam as maiores notas estando, portanto, mais alinhados às expectativas dos usuários.

CONCLUSÃO

Este trabalho envolveu duas etapas principais: 1) realização de um estudo com usuários visando identificar estratégias humanas para a sumarização multivídeo e 2) desenvolvimento da técnica HS-MVideoSumm baseada em critérios para seleção de conteúdo derivados de tal estudo.

Retomando-se o objetivo desta tese, pode-se concluir que os critérios para seleção de conteúdo derivados de estratégias humanas são capazes de produzir sumários semanticamente mais relevantes na visão do usuário final, no domínio de vídeos de notícias. Os critérios elencados no estudo foram implementados dando origem à técnica HS-MVideoSumm. Os experimentos conduzidos para avaliação subjetiva de seus resultados mostraram que os participantes atribuíram notas mais altas aos sumários multivídeo produzidos pela técnica proposta, superando trabalhos relacionados em destaque na literatura. Portanto, a abordagem adotada possibilitou mapear os julgamentos de relevância dos usuários finais e contribuir para a geração de resultados mais próximos a suas expectativas e necessidades.

7.1 Contribuições

As principais contribuições deste trabalho são: os critérios para seleção de conteúdo derivados de estratégias humanas (Capítulo 4), a técnica de sumarização multivídeo HS-MVideoSumm¹ (Capítulo 5) e o conhecimento gerado de que a solução proposta é eficaz para o aprimoramento dos resultados da área de pesquisa.

Além disso, podem ser apontadas diversas contribuições secundárias. Nesse sentido, primeiramente, pode-se citar a implementação de critérios ainda não explorados na literatura, sendo eles: introdução, conteúdo visual, subjetividade, comentários ou opiniões de jornalistas e entrevistas. As próprias soluções propostas para os módulos da técnica HS-MVideoSumm são

¹ A técnica HS-MVideoSumm está disponível em: <https://github.com/tamires/HS-MVideoSumm>

inovadoras, sendo algumas delas destacadas a seguir. Para a detecção de subjetividade, aplica-se o reconhecimento de faces para identificar segmentos candidatos, os quais são confirmados pela investigação do texto em uma abordagem que combina análise de sentimentos e detecção de adjetivos. Já no módulo de detecção de segmentos redundantes, pode-se citar a busca por correspondências nos vetores de características textuais usando a técnica OOS, usualmente empregada no domínio de vídeos. No módulo de seleção de segmentos, por sua vez, destaca-se o cálculo de histogramas de palavras visuais com pesos TF-IDF, tradicionalmente empregados no modelo *Bag of Words*. Por fim, no módulo de produção do sumário, propõe-se uma nova abordagem para a identificação da ordem cronológica dos segmentos a serem adicionados no sumário, baseada nos conjuntos de informações redundantes.

Ainda com relação à técnica HS-MVideoSumm, ressaltam-se as avaliações objetivas de seus módulos, indicando que alcançam alta acurácia e cobertura e possibilitando seu uso em outros contextos. Os experimentos subjetivos, por sua vez, mostram que os sumários produzidos superam trabalhos relacionados reportados na literatura e estão alinhados à percepção do usuário final.

Outra contribuição é a base de vídeos no domínio de notícias², composta por cinco tópicos de destaque na mídia internacional, apresentados em edições diárias dos mais importantes telejornais brasileiros, pertencentes a diferentes categorias, cada um deles contendo três vídeos relacionados. Devido à falta de bases bem estabelecidas na literatura, esse é um passo inicial e importante para a área de pesquisa.

Este trabalho resultou em duas publicações diretamente relacionadas, citadas a seguir.

- BARBIERI, T. T. S.; GOULARTE, R. Content selection criteria for news multi-video summarization based on human strategies. **International Journal on Digital Libraries**, v. 22, n. 1, p. 1-14, 2021. ISSN 1432-5012.
- BARBIERI, T. T. S.; GOULARTE, R. Investigating subjectivity criterion for multi-video summarization. In: **Proceedings of the Brazilian Symposium on Multimedia and the Web**. New York, NY, USA: Association for Computing Machinery, 2020. (WebMedia '20), p. 137-144. ISBN 9781450381963.

Aponta-se ainda como contribuição o aperfeiçoamento profissional da autora, propiciando maior experiência com métodos de pesquisa e, em especial, com o tema abordado. Além disso, a pesquisa conduzida deu origem a um trabalho de iniciação científica (concluído) e a dois trabalhos de mestrado (em andamento).

² A base de vídeos desenvolvida está disponível em: <https://github.com/tamires/base-de-videos>

7.2 Limitações e trabalhos futuros

A principal limitação deste trabalho é não incluir como critério para a sumarização multivídeo o tempo máximo para o sumário. Optou-se por não impor tal limitação para propiciar uma análise mais clara dos resultados, isto é, sem outras restrições, obtém-se um sumário com conteúdo totalmente baseado em critérios derivados de estratégias humanas, sem a influência do corte de informações por tempo. Porém, a incorporação do tamanho, bem como outras características de personalização são possibilidades de trabalhos futuros.

Além disso, devido ao escopo delimitado para este trabalho, não foi possível incorporar todos os critérios de seleção de conteúdo identificados no estudo com usuários. Logo, os itens informação única, segmentos com fonte de informação e informações complementares podem ser futuramente implementados e inseridos na técnica.

O trabalho como um todo possibilita diversos desdobramentos, configurações e/ou recursos escolhidos para o desenvolvimento de cada módulo da técnica HS-MVideoSumm podem ser substituídos e testados. Exemplos: é possível utilizar outro critério para a escolha da introdução que não o seu tamanho; os modelos *Bag of Words* para a representação do texto e *Bag of Features* para o conteúdo visual podem ser substituídos para se verificar o impacto nos conjuntos de informações redundantes e escolha de segmentos, respectivamente; a técnica de busca por correspondências nos vetores de características textuais pode dar lugar a algoritmos de agrupamento.

Outra limitação que pode ser apontada é o tamanho da base de vídeos, por seu desenvolvimento ter se iniciado neste trabalho e envolver a participação de usuários na segmentação em subtópicos. Tal base pode ser estendida em trabalhos futuros.

Além disso, a técnica HS-MVideoSumm é dependente de um serviço externo para análise de sentimentos (*Google Cloud Natural Language API*), necessitando de ajustes nos limiares definidos em caso de atualizações do serviço. Cita-se ainda que, como entrada, foram utilizados subtópicos identificados de modo manual, sendo uma possibilidade de trabalho futuro a automatização desse processo. Por fim, a técnica proposta foi avaliada no domínio de notícias, porém é possível investigar a viabilidade de aplicação em outros contextos.

REFERÊNCIAS

- AGOSTINI, V.; LÓPEZ CONDORI, R. E.; PARDO, T. A. S. Automatic alignment of news texts and their multi-document summaries: Comparison among methods. In: _____. **Computational Processing of the Portuguese Language**. New York, NY, USA: Springer International Publishing, 2014. p. 220–231. ISBN 978-3-319-09761-9. Citado na página 23.
- AIZAWA, K.; TANCHAROEN, D.; KAWASAKI, S.; YAMASAKI, T. Efficient retrieval of life log based on context and content. In: **Proceedings of the the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences**. New York, NY, USA: ACM, 2004. (CARPE'04), p. 22–31. ISBN 1-58113-932-2. Citado na página 45.
- AKRAMULLAH, S. **Digital Video Concepts, Methods, and Metrics: Quality, Compression, Performance, and Power Trade-off Analysis**. 1st. ed. New York, NY, USA: Apress, 2014. 368 p. ISBN 9781430267133. Citado nas páginas 29 e 30.
- ALMEIDA, J.; LEITE, N. J.; TORRES, R. da S. Online video summarization on compressed domain. **Journal of Visual Communication and Image Representation**, v. 24, n. 6, p. 729–738, 2013. ISSN 1047-3203. Citado na página 22.
- BARBIERI, T. T. S. **Representação de tomadas como suporte à segmentação em cenas**. 93 p. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) — Universidade de São Paulo, 2014. Citado na página 31.
- BAY, H.; TUYTELAARS, T.; VAN GOOL, L. Surf: Speeded up robust features. In: **Computer Vision – ECCV 2006**. Berlin / Heidelberg: Springer, 2006, (Lecture Notes in Computer Science, v. 3951). p. 404–417. ISBN 978-3-540-33832-1. Citado nas páginas 38 e 40.
- BLANKEN, H. M.; VRIES, A. P.; BLOK, H. E.; FENG, L. **Multimedia Retrieval**. Berlin / Heidelberg: Springer, 2007. 372 p. ISBN 3540728945. Citado nas páginas 26, 33, 34, 45, 46 e 48.
- CAMARGO, R. T.; DI-FELIPPO, A.; PARDO, T. A. S. On strategies of human multi-document summarization. In: **Proceedings of Symposium in Information and Human Language Technology**. Natal, RN, Brasil: Sociedade Brasileira de Computação, 2015. p. 141–150. Citado na página 23.
- CHANG, S.-F.; VETRO, A. Video adaptation: Concepts, technologies, and open issues. **Proceedings of the IEEE**, v. 93, n. 1, p. 148–158, Jan 2005. ISSN 0018-9219. Citado na página 42.
- CHEN, F.; DE VLEESCHOUWER, C.; CAVALLARO, A. Resource allocation for personalized video summarization. **IEEE Transactions on Multimedia**, v. 16, n. 2, p. 455–469, 2014. ISSN 1520-9210. Citado na página 22.
- COIMBRA, D. B. **Segmentação de cenas em telejornais: uma abordagem multimodal**. 107 p. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) — Universidade de São Paulo, 2011. Citado na página 30.

COMPESI, R. **Video Field Production and Editing**. 7th. ed. Oxford, UK: Focal Press, 2015. 484 p. ISBN 9781317351108. Citado nas páginas 25 e 26.

DALE, K.; SHECHTMAN, E.; AVIDAN, S.; PFISTER, H. Multi-video browsing and summarization. In: **IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops**. New York, NY, USA: IEEE, 2012. p. 1–8. ISSN 2160-7508. Citado nas páginas 53 e 54.

DIMITROVA, N.; AGNIHOTRI, L.; BARBIERI, M.; WEDA, H. Video segmentation. In: _____. **Encyclopedia of Database Systems**. Boston, MA, USA: Springer, 2009. p. 3308–3313. ISBN 978-0-387-39940-9. Citado nas páginas 22 e 30.

GROOTENDORST, M. **9 Distance Measures in Data Science**. 2021. Acessado em 26/03/2021. Disponível em: <<https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa>>. Citado na página 40.

GYGLI, M.; GRABNER, H.; RIEMENSCHNEIDER, H.; VAN GOOL, L. Creating summaries from user videos. In: **Computer Vision – ECCV 2014**. New York, NY, USA: Springer International Publishing, 2014, (Lecture Notes in Computer Science, v. 8695). p. 505–520. ISBN 978-3-319-10583-3. Citado na página 22.

HERVE, N.; BOUJEMAA, N. Automatic image annotation. In: _____. **Encyclopedia of Database Systems**. Boston, MA, USA: Springer, 2009. p. 180–187. ISBN 978-0-387-35544-3. Citado na página 86.

HONG, R.; TANG, J.; TAN, H.-K.; YAN, S.; NGO, C.; CHUA, T.-S. Event driven summarization for web videos. In: **Proceedings of the First SIGMM Workshop on Social Media**. New York, NY, USA: ACM, 2009. (WSM '09), p. 43–48. ISBN 978-1-60558-759-2. Citado nas páginas 53, 54 e 60.

HUA, X.-S.; WANG, M. Video content structure. In: _____. **Encyclopedia of Database Systems**. Boston, MA, USA: Springer, 2009. p. 3281–3286. ISBN 978-0-387-39940-9. Citado na página 31.

HUANG, X.; WANG, J.; ZHANG, M.; ZHAI, J. Gradual-surf. In: **2011 4th International Congress on Image and Signal Processing (CISP)**. New York, NY, USA: IEEE, 2011. v. 2, p. 906–909. Citado na página 40.

IDE, I.; ZHANG, Y.; TANISHIGE, R.; DOMAN, K.; KAWANISHI, Y.; DEGUCHI, D.; MURASE, H. Summarization of news videos considering the consistency of auditory and visual contents. In: **2017 IEEE International Symposium on Multimedia (ISM)**. New York, NY, USA: IEEE, 2017. p. 193–199. Citado nas páginas 21, 22, 53, 54, 60 e 99.

JI, Z.; ZHANG, Y.; PANG, Y.; LI, X. Hypergraph dominant set based multi-video summarization. **Signal Processing**, Elsevier North-Holland, Inc., New York, NY, USA, v. 148, n. C, p. 114–123, Jul 2018. ISSN 0165-1684. Citado nas páginas 21, 49, 50, 51, 54, 56, 60 e 104.

JI, Z.; ZHANG, Y.; PANG, Y.; LI, X.; PAN, J. Multi-video summarization with query-dependent weighted archetypal analysis. **Neurocomputing**, v. 332, p. 406–416, 2019. ISSN 0925-2312. Citado nas páginas 22, 23, 54, 56, 60, 102 e 104.

JI, Z.; ZHAO, Y.; PANG, Y.; LI, X. Cross-modal guidance based auto-encoder for multi-video summarization. **Pattern Recognition Letters**, v. 135, p. 131–137, 2020. ISSN 0167-8655. Citado nas páginas 22, 54, 57, 60, 102 e 104.

KANEHIRA, A.; VAN GOOL, L.; USHIKU, Y.; HARADA, T. Viewpoint-aware video summarization. In: **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**. New York, NY, USA: IEEE, 2018. p. 7435–7444. ISSN 2575-7075. Citado nas páginas 23, 54, 55 e 104.

KANNAN, R.; GHINEA, G.; SWAMINATHAN, S.; KANNAIYAN, S. Improving video summarization based on user preferences. In: **Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)**. New York, NY, USA: IEEE, 2013. p. 1–4. Citado na página 21.

KITCHENHAM, B. **Procedures for performing systematic reviews**. Keele, UK, 2004. 26 p. Citado na página 51.

LEW, M. S.; SEBE, N.; DJERABA, C.; JAIN, R. Content-based multimedia information retrieval: State of the art and challenges. **ACM Trans. Multimedia Comput. Commun. Appl.**, ACM, New York, NY, USA, v. 2, n. 1, p. 1–19, 2006. ISSN 1551-6857. Citado na página 45.

LI, P.; YE, Q.; ZHANG, L.; YUAN, L.; XU, X.; SHAO, L. Exploring global diverse attention via pairwise temporal relation for video summarization. **Pattern Recognition**, v. 111, p. 107677, 2021. ISSN 0031-3203. Citado na página 22.

LI, Y.; MERIALDO, B. Multi-video summarization based on av-mmr. In: **International Workshop on Content-Based Multimedia Indexing (CBMI)**. New York, NY, USA: IEEE, 2010. p. 1–6. ISSN 1949-3983. Citado nas páginas 54, 55 e 60.

_____. Multi-video summarization based on Video-MMR. In: **11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10**. New York, NY, USA: IEEE, 2010. p. 1–4. ISSN 2158-5873. Citado nas páginas 54, 55, 60, 102 e 103.

_____. Multi-video summarization based on ob-mmr. In: **9th International Workshop on Content-Based Multimedia Indexing (CBMI)**. New York, NY, USA: IEEE, 2011. p. 163–168. ISSN 1949-3983. Citado nas páginas 54, 55 e 60.

_____. Video summarization based on balanced av-mmr. In: _____. **Advances in Multimedia Modeling. MMM 2012. Lecture Notes in Computer Science, vol 7131**. Berlin / Heidelberg: Springer, 2012. p. 370–382. ISBN 978-3-642-27355-1. Citado nas páginas 54, 55 e 60.

_____. Multimedia maximal marginal relevance for multi-video summarization. **Multimedia Tools and Applications**, v. 75, n. 1, p. 199–220, 2016. ISSN 1573-7721. Citado nas páginas 22, 23, 30, 54, 55, 60, 102 e 103.

LI, Y.; MERIALDO, B.; ROUVIER, M.; LINARES, G. Static and dynamic video summaries. In: **Proceedings of the 19th ACM International Conference on Multimedia**. New York, NY, USA: ACM, 2011. (MM '11), p. 1573–1576. ISBN 978-1-4503-0616-4. Citado nas páginas 54, 55 e 60.

LOWE, D. Object recognition from local scale-invariant features. In: **The Proceedings of the Seventh IEEE International Conference on Computer Vision**. New York, NY, USA: IEEE, 1999. v. 2, p. 1150–1157. Citado na página 38.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. **International Journal of Computer Vision**, Springer Netherlands, v. 60, p. 91–110, 2004. ISSN 0920-5691. Citado nas páginas 38, 39 e 40.

- MAGALHÃES, J.; PEREIRA, F. Using mpeg standards for multimedia customization. **Signal Processing: Image Communication**, v. 19, n. 5, p. 437–456, 2004. ISSN 0923-5965. Citado nas páginas 42 e 43.
- MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **An Introduction to Information Retrieval**. Cambridge, UK: Cambridge University Press, 2009. Citado nas páginas 35, 38, 40, 46 e 47.
- MARKET.US. **Amazon Prime Video Statistics and Facts**. 2020. Acessado em 15/10/2020. Disponível em: <<https://market.us/statistics/online-video-and-streaming-sites/amazon-prime-video/>>. Citado na página 21.
- MARQUES FILHO, O.; VIEIRA NETO, H. **Processamento digital de imagens**. 1st. ed. Rio de Janeiro, RJ, Brasil: Brasport, 1999. 406 p. Citado na página 36.
- MARQUES, O. **Practical Image and Video Processing Using MATLAB**. New Jersey, USA: Wiley, IEEE Press, 2011. 696 p. Citado na página 37.
- MONEY, A. G.; AGIUS, H. Video summarisation: A conceptual framework and survey of the state of the art. **J. Vis. Comun. Image Represent.**, Academic Press, Inc., Orlando, FL, USA, v. 19, n. 2, p. 121–143, 2008. ISSN 1047-3203. Citado nas páginas 42, 43 e 44.
- NAPHIDE, H. R.; HUANG, T. S. A probabilistic framework for semantic video indexing, filtering, and retrieval. **IEEE Transactions on Multimedia**, v. 3, n. 1, p. 141–151, mar 2001. ISSN 1520-9210. Citado na página 42.
- NETFLIX. **2017 on Netflix - A Year in Bingeing**. 2017. Acessado em 15/10/2020. Disponível em: <<https://media.netflix.com/en/press-releases/2017-on-netflix-a-year-in-bingeing>>. Citado na página 21.
- _____. **About Us - About Netflix**. 2020. Acessado em 15/10/2020. Disponível em: <<https://about.netflix.com/en/about-us#about-netflix>>. Citado na página 21.
- NIE, L.; HONG, R.; ZHANG, L.; XIA, Y.; TAO, D.; SEBE, N. Perceptual attributes optimization for multivideo summarization. **IEEE Transactions on Cybernetics**, v. 46, n. 12, p. 2991–3003, Dec 2016. ISSN 2168-2267. Citado nas páginas 54, 58 e 99.
- OU, S. H.; LEE, C. H.; SOMAYAZULU, V. S.; CHEN, Y. K.; CHIEN, S. Y. On-line multi-view video summarization for wireless video sensor network. **IEEE Journal of Selected Topics in Signal Processing**, v. 9, n. 1, p. 165–179, Feb 2015. ISSN 1932-4553. Citado nas páginas 54, 56 e 60.
- OU, S. H.; LU, Y. C.; WANG, J. P.; CHIEN, S. Y.; LIN, S. D.; YETI, M. Y.; LEE, C. H.; GIBBONS, P. B.; SOMAYAZULU, V. S.; CHEN, Y. K. Communication-efficient multi-view keyframe extraction in distributed video sensors. In: **IEEE Visual Communications and Image Processing Conference**. New York, NY, USA: IEEE, 2014. p. 13–16. Citado nas páginas 54, 55, 60 e 102.
- PANDA, R.; MITHUN, N. C.; ROY-CHOWDHURY, A. K. Diversity-aware multi-video summarization. **IEEE Transactions on Image Processing**, v. 26, n. 10, p. 4712–4724, Oct 2017. ISSN 1057-7149. Citado nas páginas 22, 51, 54, 57, 60 e 102.

PANDA, R.; ROY-CHOWDHURY, A. K. Collaborative summarization of topic-related videos. In: **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. New York, NY, USA: IEEE, 2017. p. 4274–4283. ISSN 1063-6919. Citado nas páginas [54](#), [57](#), [60](#) e [104](#).

PETRIE, H.; BEVAN, N. The evaluation of accessibility, usability, and user experience. In: _____. Boca Raton, FL, USA: CRC Press, 2009. p. 1–30. Citado nas páginas [63](#) e [65](#).

PETROU, M.; PETROU, C. **Image Processing: The Fundamentals**. 2nd. ed. New Jersey, USA: Wiley, 2010. 818 p. Citado na página [37](#).

POTAPOV, D.; DOUZE, M.; HARCHAOUI, Z.; SCHMID, C. Category-specific video summarization. In: **Computer Vision – ECCV 2014**. New York, NY, USA: Springer International Publishing, 2014, (Lecture Notes in Computer Science, v. 8694). p. 540–555. ISBN 978-3-319-10598-7. Citado na página [22](#).

PREECE, J.; ROGERS, Y.; SHARP, H. **Interaction Design: Beyond Human-Computer Interaction**. 1st. ed. New Jersey, USA: John Wiley & Sons, 2002. 544 p. ISBN 0-471-49278-7. Citado na página [65](#).

RIBALDO, R.; CARDOSO, P. C. F.; PARDO, T. A. S. Exploring the subtopic-based relationship map strategy for multi-document summarization. **Revista de Informática Teórica e Aplicada**, Porto Alegre, RS, Brasil, v. 23, n. 1, p. 183–211, Mai 2016. Citado nas páginas [102](#) e [103](#).

RICHARDSON, I. E. **The H.264 Advanced Video Compression Standard**. 2nd. ed. New Jersey, USA: John Wiley & Sons, 2010. 346 p. ISBN 9780470516928. Citado nas páginas [26](#), [27](#), [28](#) e [29](#).

RICHARDSON, I. E. G. **Video Codec Design**. 1st. ed. New Jersey, USA: Wiley, 2002. 328 p. ISBN 0471485535. Citado na página [30](#).

RUI, Y.; HUANG, T. S.; MEHROTRA, S. Constructing table-of-content for videos. **Multimedia Systems**, Springer Berlin / Heidelberg, v. 7, n. 5, p. 359–368, 1999. ISSN 0942-4962. Citado na página [30](#).

SEARS, A.; HANSON, V. L. Representing users in accessibility research. **ACM Transactions on Accessible Computing**, ACM, New York, NY, USA, v. 4, n. 2, p. 7:1–7:6, Mar 2012. ISSN 1936-7228. Citado na página [63](#).

SHAO, J.; JIANG, D.; WANG, M.; CHEN, H.; YAO, L. Multi-video summarization using complex graph clustering and mining. **Computer Science and Information Systems**, v. 7, n. 1, p. 85–98, 2010. Citado na página [22](#).

SILVA, G.; YAMASAKI, T.; AIZAWA, K. Evaluation of video summarization for a large number of cameras in ubiquitous home. In: **Proceedings of the 13th Annual ACM International Conference on Multimedia**. New York, NY, USA: ACM, 2005. p. 820–828. Citado na página [45](#).

SMEATON, A. F. Techniques used and open challenges to the analysis, indexing and retrieval of digital video. **Information Systems**, v. 32, n. 4, p. 545 – 559, 2007. ISSN 0306-4379. Citado na página [30](#).

- SMEULDERS, A. W. M.; WORRING, M.; SANTINI, S.; GUPTA, A.; JAIN, R. Content-based image retrieval at the end of the early years. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 12, p. 1349–1380, 2000. ISSN 0162-8828. Citado na página 34.
- STEHLLING, R. O.; NASCIMENTO, M. A.; FALCÃO, A. X. A compact and efficient image retrieval approach based on border/interior pixel classification. In: **Proceedings of the Eleventh International Conference on Information and Knowledge Management**. New York, NY, USA: ACM, 2002. (CIKM '02), p. 102–109. ISBN 1-58113-492-4. Citado na página 37.
- SWAIN, M. J.; BALLARD, D. H. Color indexing. **International Journal of Computer Vision**, v. 7, n. 1, p. 11–32, 1991. ISSN 1573-1405. Citado na página 79.
- TEKALP, A. M. **Digital Video Processing**. 2nd. ed. New Jersey, USA: Prentice Hall, 2015. 624 p. ISBN 9780133991116. Citado nas páginas 25 e 27.
- TOFFLER, A. **Future Shock**. 1. ed. New York, NY, USA: Bantam, 1984. 576 p. Citado na página 21.
- TSENG, B. L.; LIN, C.-Y.; SMITH, J. R. Using mpeg-7 and mpeg-21 for personalizing video. **MultiMedia, IEEE**, v. 11, n. 1, p. 42–52, jan 2004. ISSN 1070-986X. Citado na página 45.
- VIEIRA, N. **YouTube Broadcast**. 2019. Acessado em 15/10/2020. Disponível em: <https://canaltech.com.br/eventos/youtube-brandcast-confira-tudo-o-que-rolou-na-quinta-edicao-aqui-no-br-150647/>. Citado na página 21.
- WANG, F.; MERIALDO, B. Multi-document video summarization. In: **IEEE International Conference on Multimedia and Expo**. New York, NY, USA: IEEE, 2009. p. 1326–1329. ISSN 1945-7871. Citado nas páginas 53, 54 e 60.
- WANG, M.; HONG, R.; LI, G.; ZHA, Z. J.; YAN, S.; CHUA, T. S. Event driven web video summarization by tag localization and key-shot identification. **IEEE Transactions on Multimedia**, v. 14, n. 4, p. 975–985, Aug 2012. ISSN 1520-9210. Citado nas páginas 53, 54, 59 e 60.
- WU, J.; ZHONG, S.; LIU, Y. Dynamic graph convolutional network for multi-video summarization. **Pattern Recognition**, v. 107, p. 107382, 2020. ISSN 0031-3203. Citado nas páginas 22, 23, 54, 56, 60 e 102.
- YOSHITAKA, A.; SAWADA, K. Personalized video summarization based on behavior of viewer. In: **Eighth International Conference on Signal Image Technology and Internet Based Systems (SITIS)**. New York, NY, USA: IEEE, 2012. p. 661–667. Citado na página 21.
- YU, B.; MA, W.-Y.; NAHRSTEDT, K.; ZHANG, H.-J. Video summarization based on user log enhanced link analysis. In: **Proceedings of the Eleventh ACM International Conference on Multimedia**. New York, NY, USA: ACM, 2003. (MULTIMEDIA '03), p. 382–391. ISBN 1-58113-722-2. Citado na página 45.
- ZHAI, Y.; SHAH, M. Video scene segmentation using markov chain monte carlo. **IEEE Transactions on Multimedia**, v. 8, n. 4, p. 686–697, 2006. ISSN 1520-9210. Citado na página 30.
- ZHANG, L.; GAO, Y.; HONG, R.; HU, Y.; JI, R.; DAI, Q. Probabilistic skimlets fusion for summarizing multiple consumer landmark videos. **IEEE Transactions on Multimedia**, v. 17, n. 1, p. 40–49, Jan 2015. ISSN 1520-9210. Citado nas páginas 48, 54, 58, 60 e 99.

ZHANG, L.; XIA, Y.; MAO, K.; MA, H.; SHAN, Z. An effective video summarization framework toward handheld devices. **IEEE Transactions on Industrial Electronics**, v. 62, n. 2, p. 1309–1316, Feb 2015. ISSN 0278-0046. Citado nas páginas [54](#) e [58](#).

ZHANG, Y.; MA, C.; ZHANG, J.; ZHANG, D.; LIU, Y. An interactive personalized video summarization based on sketches. In: **Proceedings of the 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry**. New York, NY, USA: ACM, 2013. (VRCAI '13), p. 249–258. ISBN 978-1-4503-2590-5. Citado na página [21](#).

ZHANG, Y.; MA, H.; ZIMMERMANN, R. Dynamic multi-video summarization of sensor-rich videos in geo-space. In: _____. **Advances in Multimedia Modeling. MMM 2013. Lecture Notes in Computer Science, vol 7732**. Berlin / Heidelberg: Springer, 2013. p. 380–390. ISBN 978-3-642-35725-1. Citado nas páginas [54](#), [57](#) e [60](#).

ZHANG, Y.; WANG, G.; SEO, B.; ZIMMERMANN, R. Multi-video summary and skim generation of sensor-rich videos in geo-space. In: **Proceedings of the 3rd Multimedia Systems Conference**. New York, NY, USA: ACM, 2012. (MMSys '12), p. 53–64. ISBN 978-1-4503-1131-1. Citado nas páginas [54](#), [57](#), [58](#) e [60](#).

ZHANG, Y.; ZHANG, L.; ZIMMERMANN, R. Aesthetics-guided summarization from multiple user generated videos. **ACM Trans. Multimedia Comput. Commun. Appl.**, ACM, New York, NY, USA, v. 11, n. 2, p. 24:1–24:23, jan 2015. ISSN 1551-6857. Citado nas páginas [49](#), [54](#), [57](#), [58](#), [60](#), [102](#) e [104](#).

ZHANG, Y.; ZIMMERMANN, R. Efficient summarization from multiple georeferenced user-generated videos. **IEEE Transactions on Multimedia**, v. 18, n. 3, p. 418–431, March 2016. ISSN 1520-9210. Citado nas páginas [49](#), [54](#), [57](#), [58](#), [60](#), [102](#) e [104](#).

ZHANG, Z.; NIE, Y.; SUN, H.; ZHANG, Q.; LAI, Q.; LI, G.; XIAO, M. Multi-view video synopsis via simultaneous object-shifting and view-switching optimization. **IEEE Transactions on Image Processing**, v. 29, p. 971–985, 2020. ISSN 1941-0042. Citado nas páginas [53](#), [54](#), [55](#) e [60](#).

ZHAO, L.; YANG, S.-Q.; FENG, B. Video scene detection using slide windows method based on temporal constrain shot similarity. In: **IEEE International Conference on Multimedia and Expo (ICME 2001)**. New York, NY, USA: IEEE, 2001. p. 1171–1174. Citado na página [30](#).

ZHAO, W.; NGO, C.; TAN, H.; WU, X. Near-duplicate keyframe identification with interest point matching and pattern learning. **IEEE Transactions on Multimedia**, v. 9, n. 5, p. 1037–1048, 2007. Citado na página [41](#).

ZHOU, X.; ZHOU, X.; CHEN, L.; BOUGUETTAYA, A.; XIAO, N.; TAYLOR, J. A. An efficient near-duplicate video shot detection method using shot-based interest points. **IEEE Transactions on Multimedia**, v. 11, n. 5, p. 879–891, 2009. Citado na página [41](#).

ZHU, J.; LIAO, S.; LI, S. Z. Multicamera joint video synopsis. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 26, n. 6, p. 1058–1069, June 2016. ISSN 1051-8215. Citado nas páginas [54](#), [58](#) e [60](#).

GLOSSÁRIO

Extração de características: processo de derivar características (do inglês, *features*), geralmente de maneira automática, dos próprios objetos multimídia.

Quadro-chave: quadro que melhor representa o conteúdo visual de um segmento de vídeo.

Segmento: conjunto de quadros consecutivos que são homogêneos de acordo com algum critério estabelecido. Seu tamanho é usualmente definido pela aplicação. No contexto da técnica de sumarização multivídeo desenvolvida, o segmento de vídeo é delimitado pelo tamanho do subtópico, contendo informações visuais, textuais e sonoras.

Subtópico: agrupamento de frases consecutivas abordando um mesmo assunto, inserido no tópico principal.

Sumarização de vídeo: processo de obter uma versão reduzida do vídeo original, denominada sumário, que preserve suas informações mais significativas.

Sumarização multivídeo: processo de produzir um sumário a partir de um conjunto de vídeos relacionados.

Taxa de amostragem temporal: também chamada **taxa de quadros**, número de amostras do sinal de vídeo capturadas em intervalos de tempo periódicos, em geral, dada em quadros por segundo.

QUESTIONÁRIOS - ESTUDO COM USUÁRIOS

A.1 Termo de Consentimento Livre e Esclarecido

1. Você está sendo convidado para participar da pesquisa “Investigação de estratégias humanas para sumarização multivídeo”. Esta pesquisa tem o intuito de observar o processo de produção de um sumário a partir de um conjunto de vídeos relacionados.
2. Em sua participação, você irá produzir um único vídeo a partir de um conjunto de entrada, de acordo com os critérios que considerar mais relevantes.
3. Após a realização das tarefas, você poderá relatar a sua experiência por meio de um questionário.
4. A sua participação neste Estudo de Caso envolve riscos mínimos (como desconforto moral, ético ou físico) e é voluntária (você não terá nenhum benefício financeiro), podendo recusar-se a participar ou retirar seu consentimento, em qualquer fase da pesquisa, sem penalização alguma.
 - a. A qualquer momento, você pode desistir de participar e retirar seu consentimento.
 - b. Sua recusa não trará nenhum prejuízo em sua relação com o pesquisador ou com a instituição.
5. A pesquisa ocorrerá em um local e horário previamente definidos. Você será assistido pelos condutores do Estudo de Caso. O recurso a ser utilizado é o computador e o tempo total não ultrapassará 2 horas.
6. Antes e durante o curso da pesquisa, você poderá solicitar esclarecimentos a respeito dos procedimentos ou qualquer outra questão relacionada ao contexto do estudo.

7. Seus dados pessoais serão confidenciais.

a. Os dados coletados no Estudo de Caso serão analisados e todos os participantes receberão os resultados de tal análise, por meio de artigo publicado ou capítulo específico da monografia de doutorado.

b. Toda e qualquer informação coletada durante o estudo será tratada como confidencial. Os dados não serão divulgados de forma a possibilitar sua identificação, para isso usaremos siglas ou números para a apresentação dos resultados.

8. Os resultados obtidos por meio desta pesquisa serão utilizados para aprimorar uma técnica automática de sumarização multivídeo, que está sendo desenvolvida por pesquisadores do ICMC – Instituto de Ciências Matemáticas e de Computação.

9. Você receberá uma cópia deste termo onde consta o contato dos pesquisadores, podendo tirar suas dúvidas sobre o projeto e sua participação, agora ou a qualquer momento. Você poderá entrar em contato com os pesquisadores no endereço Avenida Trabalhador São-carlense, 400, Centro, São Carlos - SP, pelo telefone (16) 3373-9700.

Pesquisadores

Tamires Tessarolli de Souza Barbieri

Prof. Dr. Rudinei Goularte

E-mail

tamiresbarbieri@usp.br

rudinei@icmc.usp.br

Assinatura do participante

Assinatura do pesquisador responsável

Local:

Data: / /

A.2 Questionário pré-sessão

1. Idade: _____
2. Gênero: () F () M
3. Nível de escolaridade:
() Ensino médio
() Superior
() Pós-graduação
4. Nível de conhecimento em informática:
() Básico
() Intermediário
() Avançado
5. Você costuma procurar vídeos sobre acontecimentos importantes e/ou de seu interesse em diversas fontes?
() Sim () Não
Por quê? _____
6. Em sua opinião, quais as melhores fontes de vídeos sobre notícias? Por quê?

A.3 Questionário pós-sessão

1. Quais dos critérios abaixo você utilizou para selecionar conteúdo a ser inserido no sumário?
 - ☐ Informações que se repetem em todos os vídeos
 - ☐ Informação única (mencionada em apenas um vídeo)
 - ☐ Informação mencionada por mais tempo
 - ☐ Conteúdo das imagens
 - ☐ Local
 - ☐ Pessoas/objetos específicos
 - ☐ Outro. Qual? _____
 - ☐ Informação sonora
 - ☐ Posição no vídeo (ex.: início, fim, trechos igualmente distribuídos)
 - ☐ Outro. Qual? _____
2. Qual o tempo desejável de um sumário para os vídeos analisados? _____
3. Você utilizou um vídeo como base?
 - ☐ Sim. Qual? _____
 - ☐ Não
4. Você usaria um sumário multivídeo?
 - ☐ Sim
 - ☐ Não

FORMULÁRIOS - AVALIAÇÃO COM USUÁRIOS

B.1 Detecção de subjetividade

Sumários de notícias

No contexto deste trabalho, um sumário multivídeo é um resumo gerado a partir de um conjunto de notícias relacionadas ao mesmo tema.

A seguir, você encontrará 4 temas: 1. apresentação do bebê da família real britânica; 2. encontro dos presidentes Bolsonaro e Trump; 3. rompimento da barragem de rejeitos em Brumadinho; 4. incêndio na Catedral de Notre-Dame em Paris.

Para cada tema, foram resumidos 3 vídeos de notícias obtidas dos telejornais Jornal da Band, Jornal da Record e Jornal Nacional. São apresentadas 2 opções de sumários. Pedimos que escolha o que mais lhe agrada.

[Próxima](#)

Sumários de notícias

*Obrigatório

Apresentação do bebê da família real britânica

Assista aos vídeos abaixo e responda: qual sumário você prefere? *

- ☐ Sumário 1
- ☐ Sumário 2
- ☐ Não considero as diferenças entre os vídeos relevantes

Sumário 1

Sumário 2

Voltar

Próxima

Sumários de notícias

*Obrigatório

Encontro dos presidentes Bolsonaro e Trump

Assista aos vídeos abaixo e responda: qual sumário você prefere? *

- ☐ Sumário 1
- ☐ Sumário 2
- ☐ Não considero as diferenças entre os vídeos relevantes

Sumário 1

Sumário 2

Voltar

Próxima

Sumários de notícias

*Obrigatório

Rompimento da barragem de rejeitos em Brumadinho

Assista aos vídeos abaixo e responda: qual sumário você prefere? *

- ☐ Sumário 1
- ☐ Sumário 2
- ☐ Não considero as diferenças entre os vídeos relevantes

Sumário 1

Sumário 2

Voltar

Próxima

Sumários de notícias

*Obrigatório

Incêndio na Catedral de Notre-Dame em Paris

Assista aos vídeos abaixo e responda: qual sumário você prefere? *

- ☐ Sumário 1
- ☐ Sumário 2
- ☐ Não considero as diferenças entre os vídeos relevantes

Sumário 1

Sumário 2

[Voltar](#)[Próxima](#)

Sumários de notícias

*Obrigatório

Quais foram os critérios que você utilizou para fazer a escolha entre as opções apresentadas? *

Sua resposta

Voltar

Enviar

B.2 Comparação com trabalhos relacionados

Avaliação de Sumários de Notícias

No contexto deste trabalho, um sumário multivídeo é um resumo gerado a partir de um conjunto de notícias relacionadas ao mesmo tema.

A seguir, você encontrará 5 temas: 1. apresentação do bebê da família real britânica; 2. encontro dos presidentes Bolsonaro e Trump; 3. rompimento da barragem de rejeitos em Brumadinho; 4. incêndio na Catedral de Notre-Dame em Paris; 5. relação comercial entre Google e Huawei.

Para cada tema, foram resumidos 3 vídeos de notícias obtidos de telejornais brasileiros. São apresentados 3 sumários gerados com técnicas diferentes. Pedimos que assista a cada um deles e, considerando o entendimento da notícia e a organização das informações, lhes atribua notas de 1 (ruim) a 10 (bom).

Pesquisadores: Tamires T. S. Barbieri e Rudinei Goularte

Contato: tamiresbarbieri@usp.br / rudinei@icmc.usp.br

***Obrigatório**

Nome *

Sua resposta

Idade *

Sua resposta

Próxima

Avaliação de Sumários de Notícias

*Obrigatório

Apresentação do bebê da família real britânica

Assista aos vídeos abaixo e atribua notas de 1 (ruim) a 10 (bom) a cada sumário, considerando o entendimento da notícia e a organização das informações.

Sumário 1: <https://1drv.ms/v/s!AuovfLjR6HzOho4eqf-qZEooSndt5g?e=MSut39> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Sumário 2: <https://1drv.ms/v/s!AuovfLjR6HzOho4WnjIWwY7svWgtUg?e=hugPTd> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Sumário 3: <https://1drv.ms/v/s!AuovfLjR6HzOho4XQgVAoaPEPBUNuA?e=sGK2NN> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

[Voltar](#)

[Próxima](#)

Avaliação de Sumários de Notícias

*Obrigatório

Encontro dos presidentes Bolsonaro e Trump

Assista aos vídeos abaixo e atribua notas de 1 (ruim) a 10 (bom) a cada sumário, considerando o entendimento da notícia e a organização das informações.

Sumário 1: <https://1drv.ms/v/s!AuovfLjR6HzOho4ZPfLI7Q-9NMOC6Q?e=PzZKLH> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Sumário 2: <https://1drv.ms/v/s!AuovfLjR6HzOho4fQRnCJ0O6QBoflg?e=VBoSaM> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Sumário 3: <https://1drv.ms/v/s!AuovfLjR6HzOho4aEDLjaUYjJ33JrA?e=KqQ0AH> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Voltar

Próxima

Avaliação de Sumários de Notícias

*Obrigatório

Rompimento da barragem de rejeitos em Brumadinho

Assista aos vídeos abaixo e atribua notas de 1 (ruim) a 10 (bom) a cada sumário, considerando o entendimento da notícia e a organização das informações.

Sumário 1: <https://1drv.ms/v/s!AuovfLjR6HzOho4diCcqJuPH9hxdQ?e=hEVGTc> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Sumário 2: <https://1drv.ms/v/s!AuovfLjR6HzOho4cF8RxAO9NGfITGw?e=HsVyt6> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Sumário 3: https://1drv.ms/v/s!AuovfLjR6HzOho4gMmxxTOE_Ov9zuA?e=TMi59y *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Voltar

Próxima

Avaliação de Sumários de Notícias

*Obrigatório

Incêndio na Catedral de Notre-Dame em Paris

Assista aos vídeos abaixo e atribua notas de 1 (ruim) a 10 (bom) a cada sumário, considerando o entendimento da notícia e a organização das informações.

Sumário 1: <https://1drv.ms/v/s!AuovfLjR6HzOho4hvl5ZhQrDE46evQ?e=X6g4x1> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Sumário 2: <https://1drv.ms/v/s!AuovfLjR6HzOho4jx60tiOu3mf77ig?e=f7MhNp> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Sumário 3: <https://1drv.ms/v/s!AuovfLjR6HzOho4iF6fJD0vMV7J8NA?e=FJIG0U> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Voltar

Próxima

Avaliação de Sumários de Notícias

*Obrigatório

Relação comercial entre Google e Huawei

Assista aos vídeos abaixo e atribua notas de 1 (ruim) a 10 (bom) a cada sumário, considerando o entendimento da notícia e a organização das informações.

Sumário 1: <https://1drv.ms/v/s!AuovfLjR6HzOho4m6wINfvtzIEMwLA?e=k5G1YX> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Sumário 2: <https://1drv.ms/v/s!AuovfLjR6HzOho4ki8P47kkpN38vxw?e=SnLg59> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Sumário 3: <https://1drv.ms/v/s!AuovfLjR6HzOho4l8HAqE2dG9DueFw?e=mGDhic> *

	1	2	3	4	5	6	7	8	9	10	
Ruim	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Bom

Voltar

Enviar