Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Sumarização de Vídeos a partir de Dados Fisiológicos, Atenção Visual e Unidades de Ação Facial

Sérgio Cavalcanti de Paiva

Tese submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Modelos Computacionais e Cognitivos

Herman Martins Gomes
(Orientador)

Campina Grande, Paraíba, Brasil

SUMARIZAÇÃO DE VÍDEOS A PARTIR DE DADOS FISIOLÓGICOS, ATENÇÃO VISUAL E UNIDADES DE AÇÃO FACIAL

SÉRGIO CAVALCANTI DE PAIVA

TESE APROVADA EM 09/10/2019

HERMAN MARTINS GOMES, Ph.D, UFCG
Orientador(a)

EANES TORRES PEREIRA, Dr., UFCG
Examinador(a)

JOSÉ EUSTAQUIO RANGEL DE QUEIROZ, Dr., UFCG
Examinador(a)

TIAGO PEREIRA DO NASCIMENTO, Dr., UFPB
Examinador(a)

LEONARDO CUNHA DE MIRANDA, Dr., UFRN
Examinador(a)

CAMPINA GRANDE - PB

P149s Paiva, Sérgio Cavalcanti de.
Sumarização de vídeos a partir de dados fisiológicos, atenção visual e unidades de ação facial / Sérgio Cavalcanti de Paiva. – Campina Grande, 2020.
174 f. : il.

Tese (Doutorado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2019.
"Orientação: Prof. Dr. Herman Martins Gomes".
Referências.

1. Modelos Computacionais e Cognitivos. 2. Sumarização de Vídeos. 3. Action Units da Face. 4. Galvanic Skin Response. 5. Pulsação Cardíaca e Fixação Ocular. I. Gomes, Herman Martins. II. Título.

CDU 004.891(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA SEVERINA SUELI DA SILVA OLIVEIRA CRB-15/225

Dedico este trabalho aos meus filhos, Heitor e Sofia, Luzes da minha vida e razão das minhas lutas.

À minha esposa Luzibênia, minha companheira de vida, de jornadas e de sonhos, pelo apoio incondicional em todos os momentos, principalmente naqueles de muita incerteza, muito comuns nesta odisseia chamada doutorado.

Agradecimentos

Agradeço a **Deus**, primeiramente pela maravilhosa dádiva da vida e, durante esta jornada, por ter proporcionado força, saúde e inspiração nos diversos momentos mais complicados e nas situações mais adversas.

A meu orientador, **Herman Martins Gomes**, por sua disponibilidade, mesmo em período de férias, assim como pelo incentivo, os quais foram imprescindíveis para a realização desta pesquisa. Realço o apoio abrangente prestado, a forma entusiasta, formidável e apropriada como acompanhou a produção desta pesquisa. Suas críticas construtivas, bem como as discussões e reflexões conduzidas ao longo das reuniões de acompanhamento, foram indispensáveis ao longo de todo o percurso. Não posso esquecer a sua grande contribuição para meu crescimento como investigador. Sou eternamente grato por todo o apoio.

A minha esposa **Luzibênia Leal de Oliveira**, como uma orientadora nos bastidores desta pesquisa, teria de repetir o que foi mencionado agradecimento anterior. Como minha esposa, tenho a grande satisfação de compartilhar todas as coisas que construímos juntos, todos os fracassos e sucessos e toda a alegria que temos em fazer parte da vida um do outro.

Aos frutos conquistados durante esta jornada, os gêmeos **Heitor Leal de Paiva** e **Sofia Leal de Paiva**, que abriram mão de passeios, brincadeiras e tempo de convivência para que eu me dedicasse a essa pesquisa. Peço ao Senhor misericordioso que não passem por problemas como aqueles que passaram nos primeiros anos escolares e, sobretudo, que sejam felizes, como e onde for e que sempre saibam lutar por seus sonhos.

Agradeço também àqueles que me trouxeram a este mundo, meus pais **Joaquim Moreira de Paiva** e **Francisca Cavalcanti Paiva**. Quero agradecer pelo apoio incondicional prestado, principalmente nos momentos de ausência para os gêmeos, pela compreensão e paciência demonstradas que sempre e em qualquer momento me ofereceram.

Além de Pais, tenho irmãos: **Aldenir Cavalcanti de Paiva** e **Daniel Cavalcanti de Paiva** que, na medida do possível, me apoiaram durante a trajetória desta pesquisa.

Agradeço ao professor **José Eustáquio Rangel de Queiroz**, por ceder prontamente o espaço físico onde foram realizados os experimentos de aquisição dos dados contidos nesta tese.

Ao Prof. **Ricardo Olinda**, agradeço pelas contribuições na parte estatística das minhas análises e principalmente pela sua generosidade e disponibilidade em me receber e esclarecer minhas dúvidas. Consonando com minha esposa, o senhor é um educador exemplar!

Muito obrigado aos coordenadores e secretárias do Programa de Pós-Graduação em Ciência da Computação - **COPIN**, sempre prestativos, nos atendendo com cordialidade e auxiliando na resolução de problemas.

Não esqueço o papel da UFCG ao longo de todo meu percurso, da graduação ao doutorado e, por isso, agradeço pelos recursos, pelas oportunidades, pela estrutura e pelo apoio que ofereceram em cada patamar desta jornada.

A todos que aceitaram participar dos experimentos e aqueles que contribuíram, direta ou indiretamente, com este estudo.

Agradeço, também, à UFRPE, por permitir o meu afastamento integral para a qualificação, o que foi fundamental para que eu pudesse desenvolver esta pesquisa da melhor forma possível.

Aos que não mencionei, não estão esquecidos. Se de algum modo contribuíram em minha caminhada rumo ao doutoramento, deixo-lhe um agradecimento sincero.

Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes.

Marthin Luther King

Resumo

A presente pesquisa teve como objetivo geral o desenvolvimento de uma abordagem para a sumarização de vídeos digitais, fundamentada na resposta da atenção visual (movimentos oculares), fisiológico (condutância da pele e pulsação cardíaca) e emocional (ações faciais) do telespectador. Os dados monitorados do telespectador foram adquiridos durante sessões de visualização dos vídeos. Para a consecução desse objetivo geral, fez-se necessário atingir os seguintes objetivos específicos: (1) investigar a viabilidade da sumarização automática de vídeos usando características fisiológicas, faciais e de atenção visual dos telespectadores; (2) analisar a relevância das características fisiológicas, faciais e de atenção visual dos telespectadores para o sumário automático de vídeos digitais; e (3) investigar a relação entre expressões faciais e estados emocionais induzidos pela apresentação de conteúdos multimídia. A coleta e a análise de dados aconteceu durante três conjuntos de ensaios distintos, caracterizados por diferentes coleções de vídeos, grupos de voluntários e aprimoramentos identificados como necessários após análises dos conjuntos de ensaios anteriores. Os dados do último conjunto de ensaios foram analisados conforme as seguintes etapas: sincronização e normalização, seguidas de avaliação experimental. Nas avaliações, estudou-se a viabilidade da construção de sumarizadores automáticos de vídeos a partir das características monitoradas dos telespectadores, o que sugeriu sua viabilidade, inclusive com a redução do número de características. Também foi avaliada a relação entre expressões faciais e estados emocionais dos participantes, o que forneceu evidências experimentais de tal ligação. Os resultados também evidenciaram a superioridade estatística dos modelos automáticos de sumarização afetiva em relação a uma seleção aleatória. Concluiu-se, experimentalmente, que modelos personalizados para a sumarização de vídeos podem fornecer resultados mais próximos às preferências do telespectador quando comparados a modelos genéricos.

PALAVRAS-CHAVE: Sumarização de Vídeos, *Action Units* da Face, *Galvanic Skin Response*, Pulsação Cardíaca e Fixação Ocular

Abstract

The present research had as general objective the development of an approach for the summarization of digital videos, based on the visual attention (eye movements), physiological (skin conductance and heartbeat) and emotional (facial actions) response of the viewer. The monitored data of the viewer was acquired during viewing sessions of the videos. To achieve this general objective, it was necessary to achieve the following specific objectives: (1) to investigate the feasibility of automatic video summarization using viewers' physiological, facial and visual attention characteristics; (2) to analyze the relevance of the physiological, facial and visual attention characteristics of viewers to the automatic summary of digital videos; and (3) to investigate the relationship between facial expressions and emotional states induced by the presentation of multimedia content. The data collection and analysis took place during three different sets of trials, characterized by different collections of videos, groups of volunteers and improvements identified as necessary after analysis of the previous trial sets. The data from the last set of tests were analyzed according to the following steps: synchronization and normalization, followed by experimental evaluation. In the evaluations, the feasibility of building automatic video summarizers from the monitored characteristics of the viewers was studied, which suggested their viability, including the reduction in the number of characteristics. The relationship between facial expressions and emotional states of the participants was also assessed, which provided experimental evidence of such a link. The results also showed the statistical superiority of the automatic models of affective summarization to a random selection. It was concluded, experimentally, that customized models for summarizing videos can provide results closer to the viewer's preferences when compared to generic models.

KEYWORDS: Video Summarization, Action Units, Galvanic Skin Response, Heart Rate, Eye Fixation

Conteúdo

Lista de Siglas e Abreviações

AS	*Automatic Summary* (Sumário Automático)
AU	*Action Unit* (Unidade de Ação Facial)
BVP	*Blood Volume Pulse* (Fluxo do Volume de Sangue)
CAAE	Certificado de Apresentação para Apreciação Ética
caret	*Classification And REgression Training*
CE-CLM	*Convolutional Experts Constrained Local Model*
CEP	Comitê de Ética em Pesquisas com seres humanos
CNS	Conselho Nacional de Saúde
COGNIMUSE	*Cognitive Perspectives of Multimodal Signal and Event Processing*
CUS	*Comparison of User Summaries* (Comparação de Sumários de Usuários)
DP	Dilatação Pupilar
EDA	*Electrodermal Activity* (Atividade Eletrodermal)
EDR	*Electrodermal Response* (Resposta Eletrodermal)
EEG	*Electroencephalography* (Eletroencefalograma dos Sinais Neuronais)
ELVIS	*Entertainment-Led VIdeo Summaries*
EMD	*Empirical Mode Decomposition* (Decomposição de Modo Empírico)
FABO	*Bimodal Face and Body Gesture Database*
FACS	*Facial Action Coding System*
fMRI	*Functional Magnetic Resonance Imaging* (Imageamento de Ressonância Magnética Funcional)
fps	*Frames Per Second* (Quadros por Segundo)
FO	Fixação Ocular
FX	*Facial Expressions* (Expressões Faciais)
GPS	*Global Positioning System* (Sistema de Posicionamento Global)
GSR	*Galvanic Skin Response* (Resposta Galvânica da Pele)
HOGs	*Histogram of Oriented Gradients* (Histogramas de Gradiente Orientado)
HR	*Heart Rate* (Frequência Cardíaca)
HUAC	Hospital Universitário Alcides Carneiro
IM	*Interest Meter* (Medida de Interesse)
IMF	*Intrinsic Mode Function* (Funções de Modo Intrínseco)
KNN	*K-Nearest Neighbor*
LTE	*Long Term Excitement*
MCG	Melhor Combinação Global
MOT	Melhor Opção de Treinamento
MPEG	*Moving Picture Expert Group*
MS	Ministério da Saúde
MU	*Motion-Units* (Unidade de Movimento)

PD *Pupillary Dilation* (Dilatação Pupilar)
PSD *Power Spectral Density* (Densidade Espectral de Potência)
RA *Respiration Amplitude* (Amplitude da Respiração)
ROC *Reciver Operation Characteristic*
ROSE *Random Over-Sampling Examples*
RR *Respiration Rate* (Frequência Respiratória)
RS Revisão Sistemática
SaO2 Saturação de Oxigênio
SC *Skin Conductance* (Condutância da Pele)
SMOTE *Synthetic Minority Oversampling TEchinque*
SVM *Support Vector Machine*
SVR *Support Vector Regression*
TCLE Termo de Consentimento Livre e Esclarecido
TVSum *Title-based Video Summarization*
UFCG Universidade Federal de Campina Grande
US *User Summary* (Sumário do Usuário)

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Capítulo 1

Considerações Iniciais

Neste capítulo, é apresentada uma fundamentação sobre os conceitos relacionados a esta proposta de pesquisa. O capítulo tem início com uma contextualização e formalização do problema de estudo, as quais são seguidas pela apresentação dos objetivos geral e específicos. Finalmente, argumenta-se sobre a relevância da pesquisa e é apresentada a estrutura geral deste documento.

1.1 Contextualização da Pesquisa

O volume de conteúdos multimídia manipulados diariamente está crescendo rapidamente devido ao barateamento e a onipresença dos sensores, rápido desenvolvimento de técnicas de rede e plataformas de compartilhamento e o crescimento do uso das mídias sociais (ZHANG et al., 2016; YIN; THAPLIYA; ZIMMERMANN, 2018). Dentre estes tipos de conteúdo se destacam vídeos e fotos. Somente a plataforma de compartilhamento de vídeos YouTube apresenta números como um bilhão de horas assistidas diariamente[1] e quinhentas horas de vídeos são enviadas por minuto[2]. Por outro lado, mais de cem milhões de fotos e vídeos são enviados por dia no Instagram[3].

O vídeo é uma modalidade de dados multimídia com redundância inerente. Usuários interessados no conteúdo dos vídeos digitais frequentemente se deparam com a dificuldade de encontrar e consumir apenas as partes dos vídeos que lhes sejam mais úteis (HE et al., 2019).

[1] https://www.youtube.com/yt/about/press/
[2] https://www.omnicoreagency.com/youtube-statistics/
[3] https://www.omnicoreagency.com/instagram-statistics/

Técnicas de gerenciamento para armazenamento, indexação e recuperação de vídeos surgem como ferramentas indispensáveis diante do aumento quantitativo de vídeos atualmente, afirmam diversos autores (ZHANG; TAO; WANG, 2017; ZHAO; LI; LU, 2018). Dentre todas as estratégias destinadas ao processamento de vídeos em larga escala, o processo de sumarização é um passo importante para fins de recuperação, gerenciamento, assimilação e apreciação por usuários dos conteúdos mais relevantes de vídeos digitais (SAQIB; KAZMI, 2018).

Conforme Ji et al. (2019) e Paul e Salehin (2019), o processo de sumarização de vídeo consiste em extrair e oferecer ao telespectador a essência do vídeo no menor tempo possível. Tal versão sucinta do vídeo pode ser integrada a várias aplicações, tais como sistemas de busca e navegação interativa de vídeos. As técnicas de sumarização de vídeos mais difundidas na atualidade analisam o conteúdo subjacente do fluxo de vídeo para produzir uma nova versão com conteúdo reduzido a simplesmente uma lista de quadros-chave ou a uma versão de menor duração, sempre buscando refletir o mesmo conteúdo semântico do vídeo original (EJAZ et al., 2018; KOUTRAS; ZLATINSI; MARAGOS, 2018).

Para Money e Agius (2009), as técnicas de sumarização de vídeos podem ser classificadas em três categorias, a saber: **internas**, **externas** e **hibridas**, conforme explicado a seguir:

- Técnicas que utilizam informações oriundas diretamente das características da imagem, som e texto do conteúdo de vídeo, estas contidas no fluxo do vídeo, são denominadas de técnicas de sumarização interna de vídeos (por exemplo: Fu, Tai e Chen (2019), Huang e Wang (2019) e Zhang et al. (2019)).;

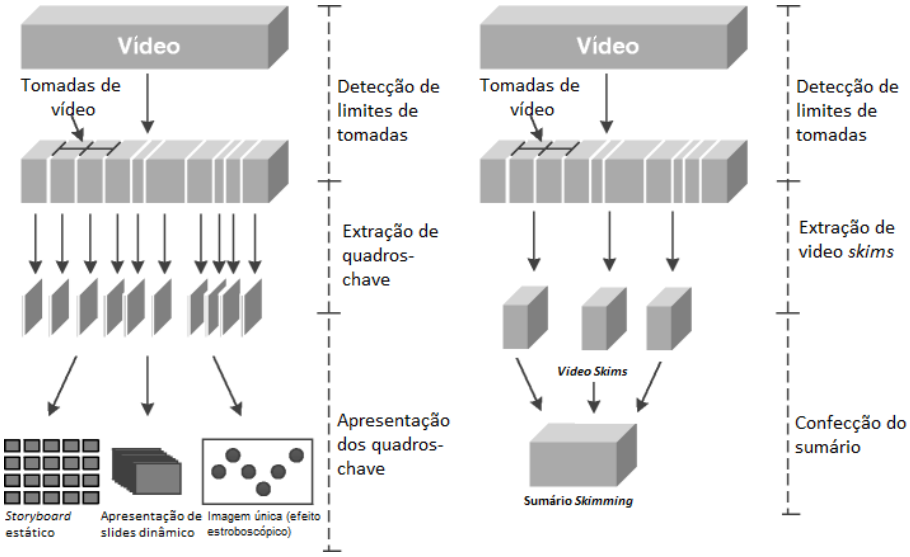
- Por outro lado, as técnicas de sumarização externa de vídeos utilizam informações coletadas externamente ao fluxo do vídeo, que podem ser baseadas no usuário ou no ambiente (por exemplo: Farouk, Dahshan e Abozeid (2016) e Fião et al. (2016)), como informações obtidas dos telespectadores e informações obtidas por exemplo de um Sistema de Posicionamento Global (*Global Positioning System* - GPS).; e

- Por fim, na intersecção das técnicas descritas anteriormente, temos as técnicas de sumarização hibrida, que analisa as informações tanto internas como externas. Uma visão geral das técnicas de sumarização pode ser encontradas em pesquisas como Truong e Venkatesh (2007), Money e Agius (2008), Sebastian e J. (2015) e K., Sen e

Raman (2019).

Na Figura 1.1, ilustram-se os esquemas de sumarização propostos por Money e Agius (2009), os quais estão relacionados à forma de apresentação. O primeiro esquema (*Storyboard*) é realizada a detecção do limite da tomada para determinar as tomadas do vídeo sobre as quais a técnica escolhida de sumarização irá atuar com o objetivo de extrair os quadros-chave mais relevantes. Por outro lado, o segundo esquema (*Skimming*) é composto de um conjunto de sequências de vídeo (*skims*) que representam os segmentos de vídeo mais pertinentes ao vídeo original. Ambos os esquemas preservam a ordem temporal do vídeo original.

Figura 1.1: Esquemas de sumarização de vídeo relacionados à forma de sua apresentação



Sumário de vídeo *Storyboard*

Fonte: **Ferreira, Cruz e Assunção (2016)**

Sumário de vídeo *Skimming*

Fonte: **Autor**

Nos últimos anos, tem havido um crescente interesse em sumarização externa de vídeos, mais especificamente na sumarização baseada na percepção do telespectador. Este tipo de sumarização utiliza conceitos de alto nível que representam a forma como os usuários podem perceber o conteúdo de um vídeo (MONEY; AGIUS, 2008). Esta abordagem apresenta a vantagem de buscar reduzir o problema do *gap* semântico entre as características multimídia de baixo nível extraídas do vídeo e conceitos de alto nível percebidos pelos telespectadores, bem como de tornar viável a criação de sumários personalizados.

O desenvolvimento de novas tecnologias e o barateamento dos custos envolvidos para o registro das informações externas do telespectador, enquanto este assiste aos conteúdos de

vídeos (KATTI et al., 2011; SINGHAL et al., 2018), fizeram surgir o ramo da sumarização afetiva de vídeos, que considera o estado afetivo do telespectador durante a apresentação de um vídeo para a geração do sumário. Assim, a partir desse tipo de sumarização, os vídeos são sintetizados com base em uma análise das informações atinentes ao estado afetivo "observado" nos telespectadores. Tais informações são classificadas como externas ao vídeo.

A sumarização de vídeos a partir da análise afetiva do telespectador possui as vantagens supracitadas, porém podem existir desvantagens relacionadas à aquisição dos dados e posterior geração dos sumários, as quais pode-se destacar: A aquisição dos dados pode requerer que o telespectador fique conectado a sensores durante o monitoramento de seus sinais fisiológicos ou que se mantenha em certa postura corporal e/ou distância de câmeras durante a análise da postura ou da face. Além disso, para a geração dos sumários são necessárias aquisições prévias em que o telespectador assista a vídeos, para fins de calibração/treinamento do sistema sumarizador.

Diante do exposto, esta pesquisa realizou o levantamento de informações relacionadas ao estado fisiológico e expressões faciais do telespectador durante a exibição de um conjunto de vídeos. Para tanto, foi desenvolvida uma abordagem que estabeleceu uma relação entre as características do telespectador e os trechos dos vídeos que eles consideraram mais relevantes.

A pesquisa ora reportada foi norteada pela seguinte questão de pesquisa: **É possível conceber uma abordagem de sumarização de vídeos baseada, de forma personalizada, em indicadores de emoções humanas?**

Mais especificamente, de posse dos dados de estado fisiológico, de atenção visual e unidades de ação facial de um participante, durante a exibição de um vídeo e sabendo o número exato de quadros que devem pertencer a um sumário, buscou-se avaliar se é possível prever os quadros que este participante selecionará após a exibição deste vídeo.

1.2 Objetivos

O objetivo geral foi desenvolver uma abordagem de sumarização de vídeos relacionada com o estado fisiológico e com as expressões faciais do telespectador monitoradas durante a exibição de vídeos.

Neste sentido, para a concepção da abordagem, foi necessário atingir os seguintes objetivos específicos:

- Investigar a viabilidade da sumarização automática de vídeos usando características fisiológicas, faciais e de atenção visual dos telespectadores;
- Analisar a relevância das características fisiológicas, faciais e de atenção visual dos telespectadores para o sumário automático de vídeos digitais; e
- Investigar a relação entre expressões faciais e estados emocionais induzidos pela apresentação de conteúdos multimídia.

1.3 Relevância

São notórios, os fatos de que técnicas eficientes de navegação são necessárias frente à quantidade de vídeos digitais disponíveis atualmente, assim como que a navegação necessita de uma apresentação resumida que seja representativa desse conteúdo. Assim, os processos de sumarização de vídeos estão se tornando ferramentas importantes para um consumo mais eficiente dos vídeos (FENG et al., 2018). Para vídeos longos, ao serem removidas as partes menos relevantes, produz-se um sumário contendo apenas as partes mais relevantes, objetivando-se uma duração inferior àquela do vídeo original.

Pesquisas realizadas na área de percepção humana indicam que o processo de sumarização de vídeos baseado em percepção permite fornecer sequências dos vídeos mais próximas daquelas esperadas pelo telespectador (MONEY; AGIUS, 2008; MEHMOOD et al., 2016). Esse processo de sumarização encontra as sequências que são mais relevantes a um telespectador ou um conjunto de telespectadores, objetivando a criação de um sumário. Por outro lado, os processos de sumarização utilizando exclusivamente informações internas dos vídeos exibidos não possibilitam esse nível de sensibilidade.

1.4 Estrutura da Tese

O presente documento é composto por um total de cinco capítulos. No Capítulo 2, é apresentado um panorama das técnicas de sumarização afetiva em vídeos, a partir de uma revisão

sistemática de pesquisas relevantes da área. Descrevem-se, também, técnicas adotadas na tarefa específica de sumarização afetiva em vídeos, estratégias comumente empregadas para resolver o problema objeto de estudo, métricas de avaliação da qualidade das sumarizações e instrumentos utilizados para realização de tal tarefa.

No Capítulo 3, são descritos os procedimentos metodológicos utilizados para a sistematização desta tese, o tipo de pesquisa, o detalhamento da população e amostra pertinentes ao estudo, os instrumentos utilizados para a coleta de dados, os procedimentos adotados para a coleta dos dados, a análise dos dados, as estruturas das avaliações e os aspectos éticos relacionados a esta pesquisa.

No Capítulo 4, encontra-se o detalhamento da coleta dos três conjuntos de ensaios para a obtenção da base de dados fisiológicos, de atenção visual e unidades de ação facial dos telespectadores durante a exibição dos vídeos e investigações de alguns processos de sumarização personalizada de vídeos a partir de dados fisiológicos, de Atenção Visual e Unidades de Ação Facial. As considerações finais são retratadas no Capítulo 5. Por último, nos elementos pós-textuais, são relacionadas as referências utilizadas para o embasamento desta pesquisa, seguidas pelos anexos e apêndices.

Capítulo 2

Revisão Bibliográfica

No presente capítulo, apresentam-se as abordagens atuais que investigam ou propõem métodos que ajudem na concepção/desenvolvimento de uma solução para a temática de sumarização afetiva de vídeos a partir de dados monitorados do telespectador, assim como apresenta também as metodologias normalmente utilizadas para avaliação dos sumários. A revisão bibliográfica apresentada neste capítulo fundamentou-se em uma revisão sistemática (RS) da literatura na área (ver Apêndice A), acrescida de uma busca complementar em um periódico internacional de alto impacto na área de revisões bibliográfica em ciência da computação.

A busca complementar foi realizada no periódico *ACM Computing Surveys (CSUR)*[1] com os termos de busca: *video summarization* e *"video summarization"*, retornando 153 e 4 registros, respectivamente. Posteriormente a leitura do título de todos os registros, foram selecionados dois *surveys* para leitura e análise dos trabalhos:

- No estudo de K., Sen e Raman (2019), intitulado *Video Skimming: Taxonomy and Comprehensive Survey*, foram citados os seguintes artigos considerados relevantes a esta pesquisa:
  - Peng et al. (2010), intitulado *A real-time user interest meter and its applications in home video summarizing*;
  - Peng et al. (2009), intitulado *A user experience model for home video summarization*; e

[1] https://csur.acm.org/

  - Yoshitaka e Sawada (2012), intitulado *Personalized video summarization based on behavior of viewer*.

- Na pesquisa de Basavarajaiah e Sharma (2019), intitulada *Survey of Compressed Domain Video Summarization Techniques*, não foi encontrado qualquer estudo novo para esta revisão.

2.1 Pesquisas Relacionadas

Na presente seção, são apresentadas as publicações eleitas para esta revisão, de forma a expor os pontos mais relevantes de cada uma, relacionando-as com o objetivo deste estudo. No Quadro 2.1, apresenta a distribuição dos artigos agrupados por sua(s) modalidade(s) de informação(ões) extraída(s) e a(s) ferramenta(s) utilizada(s) para sua obtenção.

As informações baseadas no telespectador podem ser obtidas de forma invasiva ou não por meio de vários sensores. Embora métodos não invasivos sejam geralmente preferidos, estes tendem a ser ruidosos e limitados no nível de detalhamento (MONEY; AGIUS, 2008).

Por questão de organização, nas próximas subseções, foi adotada uma subdivisão dos trabalhos pela forma de aquisição das informações dos telespectadores e esta foi subdividida em sumarização externa e em sumarização hibrida, porém alguns métodos de sumarização hibrida, cronologicamente, surgiram anteriormente aos métodos de sumarização externa.

2.1.1 Aquisição não Invasiva de Informações do Telespectador

As informações obtidas dos telespectadores de forma não invasiva utilizam normalmente dispositivos de imageamento do telespectador durante a exibição do vídeo, sendo estes discretos na obtenção dessas informações.

2.1.1.1 Sumarização Externa

Essa modalidade de sumarização se caracteriza por analisar informações que não sejam provenientes diretamente do fluxo de vídeo, porém no caso de sumarização afetiva, informações obtidas apenas da observação do telespectador.

Quadro 2.1: Artigos selecionados, modalidades e ferramentas utilizadas para extração das características dos participantes na sumarização

Artigos	**Modalidade(s)**	**Sensor(es)**
Aquisição não invasiva de informações do telespectador		
Joho et al. (2009) e Joho et al. (2011)	Expressões faciais	Webcam
Peng et al. (2009)	Deslocamentos oculares e expressões faciais	Webcam
Katti et al. (2011)	Dilatação Pupilar (PD); informações PD e fixação de olhar	Sistema de rastreamento ocular
Peng et al. (2010) e Peng et al. (2011)	Detecção de piscada, a detecção de *saccades* (deslocamento do olho) e fixação de região de interesse, detecção de movimento da cabeça e reconhecimento de expressões faciais	Webcam
Yoshitaka e Sawada (2012)	Duração de fixação ocular e operações de reprodução do vídeo	Tobii T60 e controle remoto
Dammak, Wali e Alimi (2013) e Dammak, Wali e Alimi (2015)	Posturas corporais e gestos	Webcam
Paul e Salehin (2019)	Identificação de busca suave *smooth pursuit*	Tobii eye Tracker
Aquisição invasiva de informações do telespectador		
Li et al. (2010) e Han et al. (2014)	Imageamento de ressonância magnética funcional (fMRI) do cérebro	Scanner fMRI
Money e Agius (2010) e Money e Agius (2013)	Respostas eletrodermal (EDR), amplitude da respiração (RA), frequência respiratória (RR), fluxo do volume de sangue (BVP) e frequência cardíaca (HR)	Um sensor de condutância da pele para EDR, um sensor HR/BVP e um sensor de respiração para RA e RR.
Moon et al. (2012)	Eletroencefalograma dos sinais neuronais (EEG) humanos	EEG
Mehmood et al. (2015) e Mehmood et al. (2016)	EEG humanos e sinais audiovisuais dos vídeos	EEG
Salehin e Paul (2017)	EEG humanos	EEG
Muszynski et al. (2018)	atividade eletrodermal e sinais de aceleração	Bodymedia armband sensors acoplado aos dedos.
Singhal et al. (2018)	EEG humanos	EEG
Qayyum et al. (2019)	EEG humanos	EEG

Fonte: **Autor**.

Uma abordagem para a detecção de realces pessoais em vídeos e sumarização de vídeos baseada na análise da atividade facial dos telespectadores foi utilizada nos artigos de Joho et al. (2009) e Joho et al. (2011).

No primeiro artigo, Joho et al. (2009) propuseram a criação de dois modelos, o modelo de nível pronunciado e a taxa de mudança de expressão facial. O primeiro modelo foi mo-

tivado pela observação de que certas expressões faciais são mais pronunciadas que outras. O segundo modelo foi relacionado à mudança de expressão facial de uma categoria para outra. A frequência de sua mudança foi tratada como análoga a mudança que ocorre no estado afetivo do telespectador, esse modelo contabiliza o número de quadros em que a mesma categoria continua a ser dominante em cada quadro do vídeo. Os modelos de expressões pronunciadas e o modelo de frequência de mudança foram então combinados. Para comparação, foi gerado um modelo baseado em duas características de baixo nível do vídeo. A primeira característica foi relacionada à energia global do sinal de áudio e outra característica foi a medida de similaridade entre as imagens de dois quadros do vídeo. Dessa forma, assim como no modelo anterior, foi realizada a combinação dessas características. Para a avaliação do desempenho desses modelos na tarefa de sumarização personalizada, ao final da exibição do vídeo, os participantes utilizaram uma ferramenta para selecionar as partes que mais lhe envolveram (*ground truth*). As medidas utilizadas para a avaliação/comparação das técnicas foram *precision*, *recall* e *F-score*, sendo as discussões tecidas sobre essa última medida. Os autores observaram que os modelos propostos baseados nas expressões faciais (FX) obtiveram desempenho comparável aos modelos baseados em conteúdo e que a combinação dos modelos FX aparentemente apresentou melhor performance que os modelos isoladamente.

No segundo artigo, Joho et al. (2011) desenvolveram um sistema de reconhecimento de expressões faciais em tempo real composto de um algoritmo de rastreamento de face que gera um vetor de características de movimentos de certas regiões da face, essas características alimentam um classificador de rede Bayesiana. Esses vetores foram declarados como Unidades de Movimentos (MU), estas MU são relacionadas à parte inferior ou superior da face humana e foram utilizadas como as características básicas para o esquema de classificação de expressões faciais. Para a detecção dos realces pessoais em vídeo, foram utilizados como características para análise apenas as MU, um combinado delas e as categorias de expressões faciais. Dos experimentos de Joho et al. (2009) foram herdados os vídeos e a forma de obtenção do *ground truth* dos telespectadores, sendo acrescidos mais quatro telespectadores. Na avaliação de desempenho, utilizando-se a precisão média, ficou evidente que as características de movimento para detectar realces pessoais variam significantemente entre os telespectadores e que, de modo relativo, as MU na parte superior da face humana parecem ser mais indicativas de realces pessoais que aquelas da parte inferior.

Nas pesquisas de Peng et al. (2010) e Peng et al. (2011), foram analisados o piscar dos olhos, movimentos oculares, movimentos da cabeça e expressões faciais do telespectador ao assistir um vídeo para sua edição. Dessas análises, os autores produziram um modelo de atenção e um modelo de emoção, que foram construídos para estimar a medida de interesse (IM) relacionada ao telespectador.

No primeiro estudo, Peng et al. (2010), a IM foi obtida de uma média ponderada dos modelos de atenção e da emoção, em que os pesos foram ajustados conforme a probabilidade de expressões positivas em relação as emoções neutras obtidas pelo reconhecimento de expressões faciais. A avaliação do sumário obtido da IM foi realizada comparando com um sumário obtido de seleção aleatória e aquele obtido de um usuário novato rotulando manualmente. Os participantes atribuíram escores de satisfação para cada sumário. Os resultados mostraram que o sistema proposto atingiu escores mais altos que aqueles da seleção aleatória e da seleção do usuário novato.

No segundo estudo, Peng et al. (2011), fizeram uso do conceito de lógica Fuzzy para realização da fusão dos dois modelos. A avaliação do método de sumarização foi realizada pela comparação com aqueles gerados por sumários obtidos automaticamente da seleção aleatória de tomadas, sumários gerados manualmente por um usuário novato com conhecimentos básicos de edição de vídeo e um método baseado em análise perceptiva. Os participantes avaliaram todos os sumários com um escore de satisfação. O resultado da análise revelou que o sistema proposto, na média, obteve escores mais altos que os demais.

Um método de sumarização com base nas operações de reprodução do vídeo e movimentos oculares do telespectador foi proposto em Yoshitaka e Sawada (2012). As operações *Rewind*, *Pause* e *Frame by frame replay / reverse play* do controle remoto, e a duração da fixação ocular foram utilizadas para obtenção dos sumários. Para avaliação, o participante foi convidado a preencher um questionário informando as seções que assistiu com atenção ou que considerou importantes, bem como o seu grau de atenção ou importância. As métricas utilizadas na avaliação foram *precision*, *recall* e *f-measure*. Foram realizadas duas avaliações, uma avaliação relacionada às seções mais relevantes e uma avaliação relacionada ao grau de atenção ou importância. Na primeira, a sumarização proposta foi comparada com a sumarização realizada apenas com as operações de reprodução e a sumarização baseada na fixação ocular. Na segunda, a comparação foi realizada com um sumário criado por uma

amostragem periódica com comprimento de segmento e intervalo fixo e um sumário criado pelo método de sumarização proposto com o comportamento de outro participante. Em ambos os casos, o método de sumarização proposto com os dados do participante obtiveram os melhores resultados.

Um sistema de reconhecimento emocional em tempo real foi proposto por Dammak, Wali e Alimi (2013) e Dammak, Wali e Alimi (2015) para a detecção de expressões faciais e movimentos corporais, utilizando-se o *Bimodal Face and Body Gesture Database* (FABO). O sistema utilizou o método de classificação *K-Nearest Neighbor* (KNN). Trechos do vídeo foram classificados com a base de dados FABO em duas classes relacionadas à intensidade do movimento corporal pelo sistema. A decisão tomada utilizou dois passos, o primeiro foi a detecção emocional bimodal para todos os quadros detectados e o segundo foi a utilização do valor da intensidade de movimento para encontrar a diferença entre duas emoções. A estimativa da capacidade do sistema em detectar toda a resposta afetiva produzida pelo usuário foi realizada em um experimento, no qual o sistema obteve uma taxa total de sucesso de 86%, porém este foi conduzido com apenas um usuário.

O artigo de Dammak, Wali e Alimi (2015) complementa a pesquisa anterior (DAMMAK; WALI; ALIMI, 2013), comparando o resultado dessa técnica com algumas outras técnicas nesse campo (ZAWBAA et al., 2012; JOHO et al., 2011), em que evidenciaram sua técnica, principalmente, por identificar momentos importantes do vídeo para o usuário e calcular a sua intensidade. Os autores realizaram também uma comparação com duas outras técnicas, utilizando cinco usuários e quatro filmes de esportes diferentes, no qual o esquema proposto se destacou nas seguintes três categorias: "respeitando a preferência do usuário", "agradabilidade do tempo e facilidade de compreensão" e "falta de redundância", porém não informaram como obteve essas informações dos usuários.

2.1.1.2 Sumarização Híbrida

Assim como na sumarização externa, no contexto da sumarização afetiva, a modalidade de sumarização híbrida se caracteriza por utilizar informações provenientes da observação do usuário, juntamente com informações provenientes diretamente do fluxo de vídeo, como na sumarização interna.

Um sistema híbrido baseado nas variações dos movimentos oculares e das expressões

faciais do espectador durante a exibição de um vídeo caseiro foi proposto por Peng et al. (2009). O sistema atribui valores de importância para cada quadro, referentes aos movimentos oculares e às expressões faciais do espectador, assim como para o movimento de câmera do vídeo. Para calcular a importância dos quadros foi utilizada uma média ponderada desses três valores. Os pesos utilizados variavam dependendo do tipo de expressão facial e tipo de movimento de câmera. Uma taxa de sumarização foi determinada para este processo de sumarização. Para a avaliação do experimento foram convidados dez participantes que assistiram a dois vídeos. Cada participante assistiu duas vezes cada um dos vídeos, uma para obtenção dos dados para sumarização e a outra, para rotular a parte mais importante de cada tomada do vídeo, esta foi utilizada como *ground truth* do vídeo. Três sumários para cada vídeo foram obtidos, usando apenas movimentos oculares, apenas expressões faciais e usando ambas as informações. Para cada sumário foi calculada uma taxa de correspondência, que indicou que o sumário construído utilizando ambas as informações obteve os melhores resultados.

Em Katti et al. (2011), os quadros-chave mais excitantes do vídeo foram identificados utilizando-se a resposta da dilatação pupilar (DP) e fixação ocular (FO) para a geração do *story board* afetivo. A abordagem consistiu do registro das informações de movimento ocular e da DP utilizando rastreador ocular. Os passos para a detecção do *story board* foram apenas identificar a primeira FO seguida de cada pico de excitação obtida do DP, marcar os quadros correspondentes de vídeo como quadros-chave e concatená-los para compor uma sequência. Sumários obtidos para os telespectadores foram avaliados, conforme suas preferências, quanto a cenas significativas e interessantes. A sumarização baseada em DP obteve uma captura de 40% das cenas mais significantes e interessantes. A partir de outra medida de eficiência utilizada, comparou-se a abordagem a um *ground truth* anotado por um humano, proposta por Xiang e Kankanhalli (2011). Como resultado, a abordagem proposta obteve menos falsos positivos que a abordagem que propôs a medida de eficiência.

Uma estrutura para sumarização de vídeos com base nos dados de rastreamento ocular foi proposta por Paul e Salehin (2019). Esta estrutura se diferencia das demais (por exemplo, Katti et al. (2011), Peng et al. (2011)), por utilizar a busca suave (*smooth pursuit*), estado de movimento dos olhos de um espectador seguindo um objeto em movimento no vídeo. O nível de atenção dos espectadores nos quadros, utilizados no processo de sumarização, foi

obtido da quantidade de regiões salientes e movimentos dos objetos. Os objetos mais atraentes foram descobertos por um método de previsão de saliência espacial através da construção de um mapa de saliência em torno de cada ponto de observação da busca suave, para todos os espectadores, com base nas regiões do campo visual humano. A quantidade de movimento dos objetos foi identificada pelas distâncias totais entre os pontos de vista atuais e os anteriores dos espectadores durante a busca suave e utilizada como o escore de saliência do movimento. Uma pontuação agregada de saliência para cada quadro é obtida da combinação dos mapas de saliência espacial e do movimento. Esta pontuação e uma taxa de sumarização, definida pelo espectador ou padrão ao sistema, foram utilizadas para selecionar o conjunto de quadros-chave pertencentes ao sumário. Para avaliação, três especialistas em vídeos construíram um *ground truth* dos quadros-chave para cada vídeo. As métricas utilizadas para avaliação foram *precision*, *recall* e *F-measures*. Os resultados experimentais confirmam o desempenho superior do método proposto em comparação com os métodos existentes.

2.1.2 Aquisição Invasiva de Informações do Telespectador

Os artigos expostos nesta seção referem-se a abordagem a partir das quais se extraem as informações de maneira invasiva, ou seja, o usuário está ciente da presença do sensor que está coletando suas informações.

2.1.2.1 Sumarização Externa

As informações obtidas nos artigos de sumarização afetiva mencionados nesta subseção emergem da análise de eventos imperceptíveis provenientes diretamente do usuário.

A investigação preliminar proposta por Money e Agius (2009) fundamentou-se nas respostas eletrodermal (EDR), amplitude da respiração (RA), frequência respiratória (RR), fluxo do volume de sangue (BVP) e frequência cardíaca (HR) à vídeos de uma variedade de gêneros, a saber, horror, comédia, drama, sci-fi e ação. Os resultados mostraram que o gênero **horror** despertou pronunciada resposta de EDR, RA, RR e BVP, enquanto a **comédia** apresentou baixa EDR, alta RA, RR, BVP e HR, o gênero **drama** obteve respostas fisiológicas baixas para essas medidas e os gêneros **sci-fi** e **ação** mostraram alta resposta EDR.

As medidas fisiológicas analisadas no estudo anterior, (MONEY; AGIUS, 2009), foram

utilizadas por Money e Agius (2010) para apresentar a técnica ELVIS (*Entertainment-Led VIdeo Summaries*), que foi empregada para processar e analisar dados de resposta fisiológica e identificar os subsegmentos de vídeo mais divertidos, de acordo com as respostas fisiológicas do usuário ao conteúdo de vídeo. Um conjunto de ensaios foi executado com 60 usuários assistindo a um dos três segmentos de vídeos representando conteúdos dos gêneros comédia, comédia/horror e horror. Subsequentemente, foi requisitado aos usuários, auto-reportarem um conjunto de subsegmentos do vídeo, correspondente à 30% da sua duração total, que acreditassem ser mais divertido. Assim, uma análise estatística utilizando o teste t pareado foi executada para comparar os subsegmentos obtidos pela técnica ELVIS com aqueles selecionados pelo algoritmo de seleção aleatória relacionados aos subsegmentos mais interessantes reportados pelo usuário. O resultado indicou que o ELVIS produziu subsegmentos que interceptavam de forma mais precisa com os subsegmentos selecionados pelos usuários do que aqueles obtidos pela seleção aleatória.

As modalidades extraídas nos dois estudos anteriores (MONEY; AGIUS, 2009; MONEY; AGIUS, 2010) foram utilizadas em Money e Agius (2013) para realizar novos experimentos, empregando-se a técnica ELVIS com um conjunto de ensaios maior de usuários e maior quantidade de gêneros. Os autores realizaram a comparação do resultado obtido pelo ELVIS com o resultado obtido de cada medida fisiológica isoladamente e, novamente, com a seleção aleatória de subsegmentos, verificando que o ELVIS superou todas as demais modalidades em todos os gêneros, excetuando EDR para romance e HR para ação, que não tiveram diferença estatística. Foram estimadas também medidas de agradabilidade e quantidade de informação fornecida dos resumos de 4%, 10%, 25% e 100% obtidos pelo ELVIS para cada gênero e cada individuo e comparadas com resultados de Ngo, Ma e Zhang (2005) que reportavam sumários de 10%, 25% e 100%. Da comparação dos resultados obtidos para essas medidas, o ELVIS foi superior na maioria dos casos, excetuando, na taxa de sumarização de 100%, em que este obteve resultado inferior.

Na pesquisa de Moon et al. (2012) foi proposto um sistema automático de geração de clipes usando os dados de EEG para conteúdos visualizados. Posteriormente à exibição do vídeo, à coleta e à sincronização dos dados de EEG, o sistema procura partes interessantes do vídeo relacionadas a fortes estímulos físicos gerados no usuário. Para representar o nível de interesse do usuário foi adotada a medida relacionada à emoção *Long Term Excitement*

(LTE). Dos clipes obtidos, o sistema seleciona uma quantidade pré-determinada de clipes que obtiveram os maiores valores da soma ponderada do grau de interesse e duração. Os pesos do grau de interesse e duração foram, empiricamente, ajustados para 0,8 e 0,2. O experimento foi realizado com apenas um usuário. Não foi apresentado qualquer tipo de avaliação. Foram apresentadas apenas duas figuras: um exemplo do progresso do interesse do usuário baseado no LTE durante o experimento e a interface do usuário implementada.

O método de sumarização de vídeo proposto por Salehin e Paul (2017) permitiu extrair o conteúdo afetivo utilizando os componentes de alta frequência dos sinais EEG. Os sinais EEG foram decompostos em componentes de alta e baixa frequência, conhecidos como funções de modo intrínseco (IMF), pela técnica de decomposição de modo empírico (EMD). Destes IMF, os autores verificaram que os dois primeiros forneceram melhores características discriminantes para extrair conteúdos afetivos de um vídeo quando comparados aos demais IMF. Como a força do sinal EEG indica o nível de afeto do telespectador, foi calculada a força dos sinais aplicando-se a densidade espectral da potência (PSD) nestes dois IMF para cada canal. Posteriormente, a força combinada destes IMF foi estimada para cada quadro do vídeo e gerada uma curva da atenção neuronal para um vídeo. Desta curva, o sumário do vídeo foi gerado com base na razão da sumarização padrão do sistema ou preferida pelo usuário. Os resultados experimentais revelaram que a abordagem proposta apresentou um desempenho melhor pelas métricas *precision*, *recall* e *F-measure* do que o método estado da arte de Mehmood et al. (2015).

A pesquisa de Muszynski et al. (2018) investigou as reações, atividade eletrodermal e sinais de aceleração, sincronizadas dos espectadores a realces estéticos em um contexto social durante a exibição de um filme. Os pesquisadores propuseram um sistema de detecção de realces não supervisionado com base nas informações obtidas das reações fisiológicas e comportamentais dos espectadores aos estímulos apresentados no filme. Este sistema foi composto das seguintes partes: pré-processamento de sinais e estimativa e detecção de sincronização entre espectadores com base no nível de sincronização conjunta dos espectadores. Os autores compararam diversas abordagens para estimativa da sincronização entre os múltiplos sinais de espectadores. As análises para avaliação do desempenho geral do sistema foram baseadas nas curvas ROC (*Receiver Operation Characteristic*) e nas áreas sob as curvas ROC nos níveis de sincronização conjunta dos espectadores. Analisando-se os resultados

obtidos, os autores concluíram que os realces evocam emoções relacionadas aos gêneros dos filmes e que o nível de sincronização dos sinais eletrodérmicos e de aceleração dos espectadores em ambientes sociais possibilitam detectar diferentes categorias de realces estéticos, independentes do gênero do filme. Os autores observaram que as medidas de sincronização aos pares apresentam o melhor desempenho para detectar os realces estéticos em filmes; que a combinação de diversas medidas de sincronização não causa melhoria significativa no desempenho da detecção de realces estéticos; e que as medidas de atividade eletrodérmica aparentam detectar mais realces estéticos no contexto social do que as medidas de aceleração.

Na pesquisa de Singhal et al. (2018) foi proposta uma estrutura de sumarização de vídeos com base na emoção dos usuários enquanto assistiam a vídeos analisando as atividades cerebrais por meio de sinais de eletroencefalografia (EEG). As emoções feliz, triste e neutra foram extraídas dos sinais do EEG. Uma técnica de *crowdsourcing* foi utilizada para rotular a emoção em cada segmento do vídeo utilizando os sinais de cada usuário. O sumário do tipo *skimming* foi criado dos segmentos com emoções que obtiveram votos superiores a 80%. A avaliação dos sumários foi realizada usando uma consulta on-line, em que os avaliadores avaliavam, em uma escala de 1 à 5, o sumário em comparação ao vídeo original. A avaliação média de 4 (quatro) dos vídeos foi de 3,98.

Uma estrutura de sumarização personalizada de vídeos baseada em reconhecimento de emoções humanas com base em sinais de eletroencefalografia (EEG) obtidos durante a exibição dos vídeos foi proposta por Qayyum et al. (2019). As características no domínio do tempo, de frequências e *wavelets* extraídas dos sinais de EEG foram utilizadas pelo classificador SVM na categorização das emoções do expectador. As emoções, feliz, amor, triste, raiva, surpresa e neutra, foram detectadas e classificadas nos vídeos. Os quadros onde ocorreram mudanças de emoções, excetuando-se os casos quando a mudança era para neutra, foram atribuídos como quadros-chave. Como a sumarização de vídeos é personalizada, para a avaliação foi necessária a criação de um *ground truth* preparado por cada espectador com os quadros que consideraram pertencer ao seu sumário. Dois diferentes conjuntos de métricas foram utilizados para avaliar os sumários, *precision*, *recall* e *F-measure*; e Comparação de Sumários de Usuários (*Comparison of User Summary* - CUS). Os resultados experimentais demonstraram que a estrutura proposta superou outros métodos recentes de sumarização avaliados.

2.1.2.2 Sumarização Híbrida

Conforme descrita na subseção anterior, a modalidade de sumarização híbrida utiliza informações provenientes diretamente do fluxo de vídeo, mas também aquelas informações imperceptíveis provenientes do usuário, a diferença existente aqui está apenas no fato de que essa extração de informações é realizada de forma invasiva.

Li et al. (2010) documentaram um paradigma experimental que utiliza técnicas de imageamento de ressonância magnética funcional (fMRI) do cérebro para estudar a dinâmica e as interações entre fluxos multimídias e respostas do cérebro na criação de sumários de vídeos. A idéia geral deste modelo de atenção foi combinar características de baixo nível do vídeo no modelo de atenção centrado no humano que tenha correlação máxima com as respostas fMRI do cérebro. A seleção dos quadros foi realizada com base nos valores dos quadros no modelo de atenção. Os resultados são brevemente expostos a seguir:

- Dois métodos de avaliação foram utilizados para a sumarização estática do modelo em comparação ao modelo de Ma et al. (2005). O primeiro realizou uma inspeção visual na qual o modelo obteve melhores resultados, fornecendo mais quadros relevantes. O segundo, oito telespectadores foram convidados a avaliar o quadro mais relevante para cada um dos videoclipes obtidos pelos modelos e o resultado sugeriu novamente que o modelo apresenta melhor desempenho que o primeiro; e

- Para a sumarização dinâmica, foram recrutados oito participantes, que assistiram a 16 videoclipes e avaliaram seus respectivos sumários, de acordo com a agradabilidade e quantidade de informação fornecida. Os resultados também sugeriram que esse modelo apresentou o melhor desempenho.

Ampliando o paradigma de Li et al. (2010), Han et al. (2014) apresentaram uma técnica para conceber um modelo de atenção visual fundamentado na fMRI, com a ideia subjacente de combinar características de baixo nível otimizadas sob a orientação de um pequeno número de dados de treinamento fMRI. A combinação de características de baixo nível foi otimizada por intermédio do treinamento de um modelo de regressão estatística, utilizando-se uma série de dados de treinamento. Uma vez que o modelo de regressão foi obtido, tornou-se possível prever o valor de atenção de um videoclipe de teste, dadas suas características de

baixo nível. O modelo de regressão obtido forneceu as curvas de atenção visual utilizadas no processo de sumarização. A referida técnica apresenta três componentes chaves, a saber: geração de mapa de conteúdo fundamentado no modelo Bayesiano de surpresa, geração da curva de subatenção fundamentada em um método baseado em entropia e peso ótimo por combinar curvas de subatenção calculadas utilizando-se a otimização com relação à fMRI. Foram planejados e conduzidos três experimentos para avaliar e comparar a eficácia de cada componente. O primeiro e segundo avaliaram o modelo surpresa e o método baseado em entropia, respectivamente, enquanto o último testou a estrutura proposta. Para a avaliação da qualidade dos sumários gerados foram recrutados oito indivíduos. Os sumários utilizando-se o primeiro experimento se assemelharam às abordagens de abstração de vídeo estado da arte. Os resultados mostraram o desempenho superior do último experimento em relação aos demais, assim como, do segundo em relação ao primeiro.

Um modelo de atenção híbrido, utilizado para a sumarização, obtido a partir dos fluxos de dados, sinais neuronais e características audiovisuais, foi proposto em Mehmood et al. (2015). Este modelo objetivou melhorar a percepção e ainda a compreensão de vídeos digitais. O modelo fundamentou-se na integração do nível de características e do nível de decisão. Na integração do nível de características, todas as características foram fundidas para que fosse calculado um modelo de atenção agregado e então os quadros-chave fossem extraídos utilizando esse modelo. Por outro lado, no modelo do nível de decisão, os fluxos foram processados independentemente para extrair os quadros-chave. Ao final, os resultados dos dois modelos foram então combinados na fase final e os quadros redundantes foram descartados. Para a avaliação, foram apresentados a cada um dos usuários, o vídeo original e os quadros-chave extraídos pelo modelo e por duas outras técnicas. Cada usuário também analisou a qualidade de cada sumário baseado em três métricas, quantidade de informação fornecida, prazer e classificação, nas quais o método proposto obteve valores superiores as demais técnicas em todas as métricas.

Ampliando o estudo anterior, Mehmood et al. (2016) empregaram os mesmos três tipos de fluxos de dados, geraram uma curva de atenção agregada usando um mecanismo de fusão intra e inter modalidades e, finalmente, o conteúdo afetivo em cada tomada de vídeo foi extraído. Para a avaliação do sumário obtido, cada participante gerou um sumário, selecionando os quadros mais informativos e afetivos que, posteriormente, foi utilizado como

ground truth para as comparações. A eficiência do método proposto foi baseada em dois conjuntos de métricas: (1) as métricas *precision, recall* e *f-measure* e (2) o critério de avaliação baseado em comparações de sumários com o *ground truth* do TRECVID. No conjunto (1), os sumários gerados pelo modelo de atenção agregado proposto obtiveram melhorias significativas comparados aos sumários gerados pelos modelos de atenção individuais intra e inter modalidades. Para o conjunto (2), o TRECVID definiu seus critérios de avaliação baseado em comparações de sumários com o *ground truth*. Assim, um usuário humano realizou essa comparação e quantificou cada sumário baseado em três critérios: a soma de *ground truth* incluído, a soma de redundância presente e a proporção de quadros que não deveriam aparecer no sumário. Os esquemas de sumarização utilizados, baseados em atenção não visual e visual, foram aqueles propostos por Furini et al. (2009), Avila et al. (2011) e o método proposto por Ejaz, Mehmood e Baik (2013). Após uma avaliação, com exceção de alguns casos, o método proposto geralmente superou os demais esquemas de sumarização.

No Quadro 2.2, reúnem-se algumas informações extraídas dos artigos referentes aos experimentos. Este quadro complementa o Quadro 2.1 e apresenta uma organização semelhante, apresentando por ordem de prioridade de apresentação ano de publicação, ordem alfabética crescente dos autores e grupos de pesquisa.

Na Fig. 2.1, apresenta-se a distribuição das pesquisas selecionadas para esta revisão com base nos seus respectivos anos da publicação. Desta forma, fica evidente que a sumarização afetiva de vídeos se constitui uma área de estudos recente, tendo suas primeiras publicações no ano de 2009. Por este motivo, foi apresentado com quantitativo reduzido de publicações pertinentes ao contexto desta tese.

Verifica-se, em relação aos estudos selecionados, que os anos com maiores números de pesquisas publicadas foram 2010 e 2011, ambos com três publicações. Quanto ao tipo, todos são classificados como artigos científicos e em relação ao desenho metodológico, todos são estudos experimentais.

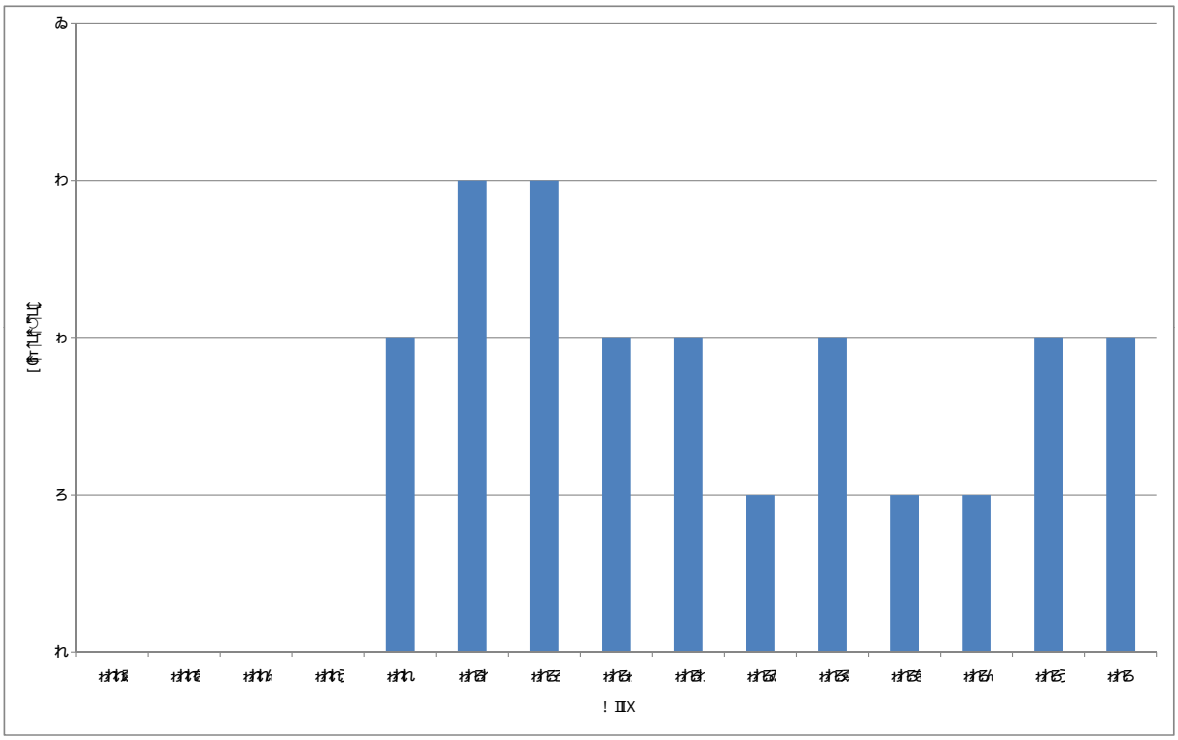
As publicações apresentaram uma distribuição variada quanto aos métodos de aquisição das informações dos telespectadores. Das vinte e uma publicações selecionadas após os critérios de inclusão e exclusão propostos neste estudo, em um total de dez, os autores utilizaram meios não invasivos para a aquisição de informações e nas demais, utilizaram meios considerados invasivos.

Quadro 2.2: Informações relevantes referentes às amostras utilizadas nos artigos

Artigos	Participantes	Vídeos			
		Quantidade	Duração	Gênero	Origem
Aquisição não invasiva de informações do telespectador					
Joho et al. (2009)	6	8 clipes	0:39.0 - 7:03.6	Variado	-
Joho et al. (2011)	10	8 clipes	0:39.0 - 7:03.6	Variado	-
Peng et al. (2009)	10	2 clipes	5 min.	-	-
Peng et al. (2010)	8	5 clipes	7 a 18 min.	-	-
Peng et al. (2011)	8	5 clipes	7 a 18 min.	-	-
Katti et al. (2011)	20	9 a 10 clipes	5 min.	Variado	YouTube
Yoshitaka e Sawada (2012)	11	1 clipe	5 min.	Futebol	-
Dammak, Wali e Alimi (2013)	1	1 clipe	5 min.	Futebol	-
Dammak, Wali e Alimi (2015)	1	1 clipe	5 min.	Futebol	-
Paul e Salehin (2019)*	4/8	4/3 clipes	-	-	-
Aquisição invasiva de informações do telespectador					
Li et al. (2010)	8	51 tomadas compondo 8 clipes	12 min.	Esportes, tempo e comercial/propaganda	TRECVID 2005
Han et al. (2014)	8	1256 vídeos	10 min. a 10 min. e 47 s	Esportes, tempo e comercial/propaganda	TRECVID 2005
Money e Agius (2010)	60 (20 por clipe)	3 clipes	35 min.	Comédia, comédia/horror e horror	-
Money e Agius (2013)	100 (20 por clipe)	5 clipes	-	Ação, drama, romance, horror e comédia	-
Moon et al. (2012)	1	1 clipe	60 min.	drama	-
Mehmood et al. (2015)	5	10 clipes	-	-	-
Mehmood et al. (2016)	10	10 clipes	-	Ação, thriller, horror e comédia	Open Video Project e AiirSource channel
Salehin e Paul (2017)*	5	3 clipes	-	Action, animação e sci-fi	YouTube
Muszynski et al. (2018)	13	30 clipes	07:22:05	Variado	-
Singhal et al. (2018)	28	4 de 20 clipes	-	-	-
Qayyum et al. (2019)	50	50 clipes	1 a 4 min. (cada)	Variado	VSUMM

Fonte: **Autor**.
Nota: (-) Informação não indicada.
* Mesmo grupo de pesquisa.

Figura 2.1: Distribuição das pesquisas por ano de publicação



Fonte: **Autor**.

Destacam-se como meios não invasivos as câmeras e os rastreadores oculares, que perfizeram um total de sete e três publicações respectivamente. No tocante aos meios invasivos, foram destacadas a utilização de EEG, sensores fisiológicos e fMRI, citados em seis, três e duas publicações respectivamente.

As bases de vídeos utilizadas são, em sua maioria, proprietárias, ou seja, não permitem a reprodutibilidade do experimento ou sua utilização em novas pesquisas. Os poucos estudos que apresentam base de vídeos de domínio público foram mencionadas em Mehmood et al. (2016) e Salehin e Paul (2017).

2.2 Metodologia de Avaliação

Novas soluções para problemas em quaisquer áreas do conhecimento necessitam avaliação objetiva quanto a eficácia e/ou eficiência de suas técnicas e de preferência contra técnicas já existentes. Na área sumarização de vídeos, os problemas de extração de quadros-chave e conjunto de sequências de vídeo são intensamente investigados, porém não existe um método melhor ou padrão para a avaliação do desempenho. O problema recai no fato da sumariza-

ção de vídeos ser um tema bastante subjetivo e, assim, impossibilitar a existência de um *ground-truth* objetivo (MA et al., 2005; YOU et al., 2007). Assim, cada pesquisa apresenta sua própria metodologia de avaliação. Segundo Truong e Venkatesh (2007), os métodos de avaliação existentes para sumários de vídeos são agrupados em três categorias: descrição do resultado, métricas objetivas e estudos de usuários.

- Descrição do resultado: é a forma de avaliação mais popular e simples, que não envolve qualquer comparação com outras técnicas;
- Métricas objetivas: para técnicas de extração de quadros-chave, uma métrica frequentemente utilizada é a função de fidelidade, calculada a partir do conjunto de quadros-chave extraído e da sequência original de quadros. A métrica é utilizada para comparar o conjunto de quadros-chave gerado por diferentes técnicas ou por uma técnica subjacente, mas com diferentes conjuntos de parâmetros; e
- Estudos de usuários: estudos que empregam usuários independentes que avaliam a qualidade dos sumários gerados e são provavelmente as formas mais úteis e realistas de avaliação.

Segundo Truong e Venkatesh (2007), as métricas chamadas objetivas são tendenciosas em direção a certos pontos de vista de sumários ou proximamente ligadas à técnica proposta. Também não existe justificativa experimental para indicar se a métrica funciona bem para o julgamento humano quanto à qualidade de um conjunto de quadros-chave.

Estudos de usuários são vistos por Song et al. (2015) como abordagens simples e rápidas, porém apresentam a desvantagem de que, caso alguma alteração seja realizada, torna-se necessário que o estudo seja executado novamente, ou seja, os sumários devem ser novamente apresentados a usuários para que sejam avaliados. Esta desvantagem é superada pelas métricas objetivas. Desde que obtidos os rótulos, os experimentos podem ser realizados indefinidamente, sendo algo desejável para sistemas de visão computacional que realizam múltiplas iterações e testes.

Diante do exposto, utilizou-se nesta pesquisa o método de avaliação proposto por Avila et al. (2011), denominado Comparação de Sumários de Usuários (*Comparison of User Summaries - CUS*), que se assemelha àquele apresentado em Guironnet et al. (2007). Neste último método, o sumário de cada vídeo é construído manualmente de quadros amostrados

em uma etapa preliminar por vários usuários (telespectadores) diferentes para o vídeo. Este sumário obtido de diversos usuários é considerado como referência, isto é, o *ground-truth*, sendo comparado com os sumários obtidos por diferentes técnicas. Porém, ao contrário do método de avaliação apresentado em Guironnet et al. (2007), o CUS realiza a comparação direta entre os sumários dos usuários e aqueles obtidos automaticamente. Dessa maneira, a opinião original de cada usuário é considerada.

O método de avaliação CUS apresenta como principais objetivos, segundo Avila et al. (2011): (i) a redução da subjetividade na tarefa de avaliação; (ii) a quantificação na qualidade do sumário; e (iii) a possibilidade de comparações entre diferentes métodos.

O método de avaliação pode ser dividido em três fases. Na primeira fase, o telespectador assiste ao vídeo e, em seguida, é convidado a construir manualmente um sumário do vídeo. Para tanto, é necessário que os quadros amostrados lhe sejam exibidos previamente. O telespectador é orientado a selecionar um conjunto de quadros que, em sua opinião, são capazes de sumarizar o conteúdo do vídeo original. Posteriormente, na segunda fase, os sumários dos telespectadores são comparados com o sumário gerado automaticamente. Na terceira fase, a qualidade do sumário gerado automaticamente é determinada por duas métricas, denominadas taxa de precisão CUS_A e taxa de erro CUS_E, definidas nas Equações 2.1 e 3.8.

$$CUS_A = \frac{n_{mAS}}{n_{US}} \quad \text{e} \tag{2.1}$$

$$CUS_E = \frac{n_{\bar{m}AS}}{n_{US}} \quad , \tag{2.2}$$

em que n_{mAS} é o número de quadros-chave correspondentes do sumário automático (AS), $n_{\bar{m}AS}$ é o número de quadros-chave não correspondentes do AS e n_{US} é o número de quadros-chave do sumário do usuário (US).

Os valores de CUS_A variam de 0 (pior caso, quando nenhum quadro-chave do AS combina com os quadros-chave do US) a 1 (melhor caso, quando todos os quadros-chave do US coincidem com os quadros-chave do AS). É necessário lembra-se de que $CUS_A = 1$ não significa necessariamente que todos os quadros-chave do AS e aqueles do US sejam compatíveis, desde que a quantidade de quadros de AS pode ser maior que a quantidade de

quadros do US. Enquanto isso, os valores de CUS_E, variam de 0 (melhor caso, quando todos os quadros-chave do AS coincidem com os quadros-chave dos US) a n_{AS}/n_{US} (pior caso, quando nenhum dos quadros do AS coincide com os quadros do US). Isto significa que as métricas CUS_A e CUS_E são complementares, possibilitando a obtenção do melhor sumário quando $CUS_A = 1$ e $CUS_E = 0$.

2.3 Conclusão da Revisão de Literatura

Esta revisão foi inicialmente idealizada para seguir um formato de revisão sistemática pelos inúmeros benefícios reportados em diversas áreas de pesquisa como Engenharia de Software, por exemplo, nas pesquisas de Catal e Diri (2009), Kitchenham et al. (2009), Breivold, Crnkovic e Larsson (2012) e Hall et al. (2012). Porém, posteriormente a uma primeira versão desta revisão, foram identificados alguns problemas. Dentre os problemas encontrados, estavam artigos pertencentes a uma base de dados bibliográfica e não encontrados nas buscas e durante as buscas nas bases de dados bibliográficas foi observado o retorno de um quantitativo considerável de artigos duplicados ou não relacionados à pesquisa. Um aspecto que poderia ser realizado em revisões futuras seria colocar um filtro de qualidade, relacionado ao fator de impacto nos veículos de publicação, assim evitando a inclusão de referências pouco representativas da área.

Diante do objetivo desta revisão, ficou evidente a carência de literatura abordando a temática - sumarização afetiva de vídeos - uma vez que só se registraram publicações envolvendo este assunto há cerca de uma década. Neste sentido, considera-se relevante a divulgação de pesquisas desta nova e promissora área do conhecimento, a qual constitui uma importante estratégia capaz de facilitar a vida moderna, por acumular diversas e numerosas informações oferecidas pelos recursos tecnológicos.

A sumarização baseada na percepção do telespectador realizada por meios discretos de aquisição como, por exemplo, com auxilio de câmeras, têm sido mais utilizados em função da facilidade de serem implantadas, devido a sua disseminação em diversos locais, enquanto que a aquisição por meios invasivos apresentam a desvantagem inerente ao contato, porém com o avanço das tecnologias, todos os sensores descritos nesta revisão poderão ser dispostos em nossos vestuários e acessórios em um futuro próximo.

Cabe destacar o número reduzido de participantes e/ou de vídeos descritos nas pesquisas que compuseram esta revisão. Em poucos experimentos, o número não ultrapassou dez usuários ou oito vídeos.

A abordagem proposta nesta tese tem como diferenciais a possibilidade de geração de sumários personalizados relacionados aos telespectadores monitorados e, ainda, a utilização conjunta de dados fisiológicos (atividade eletrodermal e batimentos cardíacos), unidades de ativação (*Action Unit*) da face e rastreamento ocular para a geração do modelo de estado emocional do telespectador. Dessa forma, o processo de sumarização nesta tese explorou, de maneira conjunta, algumas das informações que foram utilizadas separadamente em estudos anteriores, tais como Money e Agius (2009), Joho et al. (2009) e Katti et al. (2011). Vale também ressaltar que as comparações apresentadas nas avaliações desta tese foram testadas estatisticamente.

Uma comparação entre a abordagem proposta e estudos relacionados está presente no Quadro 2.3. As semelhanças foram apresentadas na Metodologia utilizada para avaliação da tarefa de sumarização (Seção 2.2) e qual(is) tipo(s) de sumário(s) foi(oram) utilizado(s) na comparação dos estudos. Como podem ser observadas muitas pesquisas utilizaram para comparação dos sumários, mais de um tipo de sumário.

Quadro 2.3: Comparação das principais pesquisas com a metodologia proposta nesta pesquisa

Características	**Pesquisas Relacionadas**	**Metodologia Proposta**
Metodologia de Avaliação na Tarefa de Sumarização		
Descrição do resultado	Dammak, Wali e Alimi (2013), Moon et al. (2012)	-
Métricas objetivas	Joho et al. (2009), Peng et al. (2009), Money e Agius (2010), Joho et al. (2011), Katti et al. (2011), Yoshitaka e Sawada (2012), Money e Agius (2013), Mehmood et al. (2016), Salehin e Paul (2017), Muszynski et al. (2018), , Paul e Salehin (2019), Qayyum et al. (2019)	X
Estudos com avaliação subjetiva de usuários	Peng et al. (2010), Peng et al. (2011), Li et al. (2010), Money e Agius (2013), Han et al. (2014), Dammak, Wali e Alimi (2015), Mehmood et al. (2015), Singhal et al. (2018)	-
Comparação dos Sumários		
Seleção aleatória	Money e Agius (2010), Peng et al. (2010), Peng et al. (2011), Money e Agius (2013)	X
Comparação com outras pesquisas	Joho et al. (2009), Li et al. (2010), Peng et al. (2011), Money e Agius (2013), Han et al. (2014), Mehmood et al. (2015), Mehmood et al. (2016), Salehin e Paul (2017), Qayyum et al. (2019)	-
Seleção de características	Peng et al. (2009), Joho et al. (2011), Yoshitaka e Sawada (2012), Money e Agius (2013), Han et al. (2014), Mehmood et al. (2016), Muszynski et al. (2018)	X
Manualmente*	Peng et al. (2010), Li et al. (2010), Katti et al. (2011), Peng et al. (2011), Paul e Salehin (2019)	-
Outras formas	Yoshitaka e Sawada (2012)	-
Não informou	Dammak, Wali e Alimi (2015)	-
Não realizou comparação	Li et al. (2010), Singhal et al. (2018)	-

Fonte: **Autor**.
Nota: (-) Não utilizado.
* Sumário obtido manualmente por outra(s) pessoa(s) que não seja(m) o participante

Outros comparativos podem ser traçados com relação a quantidade e a soma da duração dos vídeos apresentados aos participantes, e a quantidade de participantes, que são compatíveis e até superiores aos estudos, conforme pode ser observado no Quadro 2.2, da página 21.

Capítulo 3

Materiais e Métodos

Neste capítulo, são apresentados os materiais e métodos propostos para a solução do problema de sumarização de vídeos a partir da análise de dados fisiológicos, de atenção visual e unidades de ação facial do telespectador. Considerando-se as pesquisas discutidas no capítulo anterior, foi proposta uma nova abordagem baseada em estratégias e técnicas recentes de sumarização estado da arte. A abordagem adotada para construção da metodologia foi incremental, considerando a evolução dos diversos ensaios experimentais. Os detalhes da abordagem e das técnicas são apresentados nas próximas seções.

3.1 Tipo de Pesquisa

Esta pesquisa realizada se classifica como exploratória. De acordo com Wazlawick (2014), na pesquisa exploratória examina-se um conjunto de fenômenos buscando-se lacunas/situações pouco conhecidas até então. Dessa forma, constrói-se uma base para pesquisas mais elaboradas futuramente.

Quanto aos procedimentos técnicos, esta pesquisa se classifica como experimental. Conforme Gil (2017), este tipo de pesquisa é utilizado quando se determina um objeto de estudo, selecionam-se as variáveis que seriam capazes de influenciá-los, definem-se as formas de controle e de observação dos efeitos que a variável produz no objeto.

3.2 População e Amostra

No tipo de estudo realizado, não existe forma de escolher uma população ideal e, consequentemente, é difícil selecionar uma amostra com o rigor exigido pela estatística. Porém, a literatura pertinente revela a delimitação de um público-alvo, cita-se como exemplo o Estudo IMS Video in LATAM[1], envolvendo países da América Latina e realizado no ano de 2015, em que se observou que a faixa etária mais representativa de telespectadores de vídeos digitais foi aquela entre 15 e 34 anos, perfazendo um total de 56% da amostra. No tocante ao gênero, na referida pesquisa tanto homens quanto mulheres representaram 50% da amostra, cada contingente.

Com base no exposto, trabalhou-se com a amostragem não-probabilística do tipo intencional ou por julgamento. De acordo com Prodanov e Freitas (2013), as amostras intencionais consistem em um subgrupo da população que, com base nas informações disponíveis, possam ser consideradas representativas de toda a população.

Neste sentido, a amostra foi composta por estudantes da graduação em Ciência da Computação da Universidade Federal de Campina Grande (UFCG), que atenderam aos seguintes critérios de inclusão:

a) Aceitar participar voluntariamente da pesquisa e assinar o Termo de Consentimento Livre e Esclarecido (TCLE) (ANEXO I.1); e

b) Estar na faixa etária entre 18 e 34 anos.

3.3 Análise de Conteúdo Afetivo de Vídeo

Considerando a emoção como importante componente na classificação e recuperação de vídeos, surge a área de pesquisa Análise de Conteúdo Afetivo de Vídeo. Essa área pode ser dividida duas, segundo Wang e Ji (2015), conforme a abordagem utilizada: direta e implícita. A abordagem direta extrai conteúdos afetivos dos vídeos diretamente de características audiovisuais relacionadas. Por outro lado, a abordagem implícita detecta conteúdos afetivos dos vídeos baseados em uma análise automática de respostas espontâneas registradas do

[1]Disponível em: http://insights.imscorporate.com/files-web/IMS_Video_in_LatAm_study_PT.pdf

telespectador enquanto assiste aos vídeos.

A análise de conteúdo afetivo de vídeo consiste do conteúdo de vídeo, resposta não-verbais dos usuários, descritores emocionais e seus relacionamentos. O conteúdo do vídeo são todas as informações audiovisuais dele extraídas. As respostas não-verbais dos usuários incluem respostas fisiológicas e comportamentais visuais do usuário. Os descritores emocionais possibilitam a avaliação subjetiva dos usuários no tocante ao conteúdo afetivo do vídeo (WANG; JI, 2015).

3.3.1 Descritores Emocionais

A fim de se descrever as emoções em uma análise de conteúdo afetivo de vídeo deve-se, primeiramente, decidir qual abordagem adotar. Dentre as várias abordagens disponíveis, as mais utilizadas para a análise afetiva automática são as abordagens dimensional e categórica. Na abordagem categórica (ou discreta), a emoção é rotulada em diversas categorias. Na abordagem dimensional, a emoção é dividida em um espaço contínuo (CELIK, 2017). Logo, as abordagens internamente apresentam algumas variações:

- Para a abordagem discreta, diferentes conjuntos categóricos de emoções são propostas na literatura (CELIK, 2017). O uso mais frequente é o conjunto básico de categorias de Ekman (1992), a partir do qual a emoção é rotulada em seis categorias, que são felicidade, tristeza, surpresa, nojo, raiva e medo.

- Na abordagem dimensional, a maioria concorda que descrever uma resposta subjetiva utilizando três dimensões seja suficiente para sua representação. Entretanto, não existe consenso sobre os rótulos das dimensões (WANG; JI, 2015). Um dos possíveis conjuntos é aquele representado com valência, excitação e controle (dominância). Valência é caracterizada pelas respostas continuas variando de agradável à desagradável, já excitação representa a intensidade da emoção, variando de excitado à calmo e a terceira dimensão é utilizada para distinguir aquelas emoções que apresentam as duas dimensões anteriores com valores similares (por exemplo, diferenciar entre "tristeza"e "raiva"), essa dimensão varia de forma continua do sem controle à controle total (HANJALIC, 2004).

Seguindo essas teorias, dois tipos de descritores emocionais têm sido utilizados para registrar o conteúdo afetivo de um vídeo, categóricos e dimensionais. As duas abordagens têm suas limitações, já que as emoções são complexas e subjetivas. Por um lado, um número restrito de categorias discretas pode não representar a sutileza e complexidade dos estados afetivos e, por outro, utilizar escores contínuos e absolutos pode não ser suficientemente significativo devido à falta de padronizações para avaliação de emoções subjetivas (WANG; JI, 2015).

3.3.2 Respostas Fisiológicas e Respostas Comportamentais Visuais do Usuário

As respostas fisiológicas e respostas comportamentais visuais do usuário são os principais meios de obtenção de dados para a análise implícita de conteúdo afetivo de vídeo. O sistema nervoso simpático (SNS) controla os sinais fisiológicos que geram mudanças inconscientes no corpo segundo Jang et al. (2015). Em contrapartida, as expressões faciais podem ser produzidas de forma conscientes ou inconscientes. Dessa forma, os comportamentos faciais fornecem aparentemente, informações menos fieis à emoção que os sinais fisiológicos. Porém as formas de obtenção dos sinais fisiológicos normalmente são relacionados a instalação de sensores próximos ao corpo, enquanto as expressões faciais necessitam apenas de uma câmera para obtenção de seus registros. Assim, o comportamento visual espontâneo é mais conveniente e menos invasivo para realizar suas medições que os sinais fisiológicos, embora seus dados estejam sujeitos a problemas relacionados com a resolução da câmera e condições de iluminação.

3.3.2.1 Respostas Fisiológicas

A pulsação cardíaca é entendida como a contração exercida pelos ventrículos do coração bombeando o sangue para dentro das artérias, conforme Craven e Hirnle (2006) e Atkinson e Murray (2008). A pressão do sangue entrando na aorta a partir do ventrículo esquerdo causa o estiramento ou distensão da parede aórtica elástica. A movimentação da aorta, primeiro se expande para, em seguida, contrair-se, criando uma onda de pulso que segue ao longo dos vasos sanguíneos. A pulsação ou onda de pulso se torna perceptível como uma ascensão ou

elevação nas artérias que se apresentam próximas à superfície da pele.

De acordo com Craven e Hirnle (2006), são considerados dentro da faixa de normalidade valores entre 60 e 100 pulsações por minuto, sendo considerado bradicardia as medições inferiores a 60 e taquicardia quando ultrapassam 100 pulsações. Alguns fatores aumentam a frequência do pulso, a exemplo das emoções (medo, excitação, angustia e alegria).

Segundo Westerink et al. (2008), Maaoui e Pruski (2010) e Jang et al. (2015), a excitação do sistema nervoso autônomo estimula as glândulas sudoríparas a produzirem mais suor, aumentando assim a condutividade da pele, que pode ser medida pela resposta galvânica da pele (*Galvanic skin response* - GSR), também conhecida como atividade eletrodermal (*electrodermal activity* - EDA), condutância da pele (*skin conductance* - SC) e resposta eletrodermal (*electrodermal response* - EDR). Desta forma, as mudanças no nível de umidade da pele (suor) variam sua condutância da pele e podem revelar mudanças no sistema nervoso simpático, que é um bom indicador do nível de excitação de um indivíduo devido a estímulos sensoriais e cognitivos externos.

O corpo humano apresenta dois tipos de glândulas sudoríparas, a apócrina e a écrina. Esta última tem como função principal a termorregulação, porém se acredita que aqueles localizados nas superfícies palmar e plantar estejam mais relacionados a estímulos psicológicos do que a estímulos térmicos. Os ductos de suor na atividade eletrodermal podem ser relacionados a conjuntos de resistores variáveis ligados em paralelo, nos quais quanto mais alto o suor, menor a sua resistência (CACIOPPO; TASSINARY; BERNTSON, 2007).

A condutância da pele é registrada usando-se dois eletrodos, ambos colocados em locais ativos. As posições mais comuns dos eletrodos são as eminências tenares das palmas das mãos e a superfície volar das falanges mediais ou distais dos dedos (NOURBAKHSH et al., 2012).

Para Pruitt e Jacobs (2004), uma leitura de oxigênio no sangue indica o percentual de moléculas de hemoglobina no sangue arterial que estão saturadas com oxigênio. A leitura pode ser referida com SaO2. Leituras normais em um adulto saudável variam de 95% a 100%.

A temperatura da pele mede a resposta termal da pele humana e suas variações são normalmente relacionadas a mudanças localizadas no fluxo sanguíneo causadas pela pressão arterial ou resistência vascular. A resistência vascular local é regulada pelo tônus do mús-

culo liso, o qual é controlado pelo sistema nervoso simpático. A forma que ocorre a variação da pressão arterial pode ser definida com um modelo de regulação cardiovascular pelo sistema nervoso autônomo. Assim, a variação de temperatura da pele tem relação direta com a atividade do sistema nervoso autônomo (MAAOUI; PRUSKI, 2010) e (JANG et al., 2015). De acordo com Borg (2012), as emoções negativas como raiva, tristeza, aversão e medo desencadeiam reações fisiológicas significativas, por exemplo, o aumento da temperatura da pele e frequência cardíaca.

Também se faz necessário destacar que a temperatura corporal é uma medida muito estável, que sofre discretas variações ao longo de um dia inteiro. Neste sentido, Raff e Levitzky (2012) inferem que os seres humanos são endotérmicos, isto é, capazes de produzir seu próprio calor interno e também são considerados homeotérmicos, ou seja, mantém a temperatura corporal dentro de limites estreitos, apesar das grandes variações da temperatura ambiental. Cabe destacar a existência de um ritmo diário (circadiano) da temperatura corporal, com o valor mais baixo no inicio da manhã e o valor mais alto ao entardecer. A temperatura ainda pode sofrer variações, dependendo da atividade muscular e do ciclo menstrual, no caso das mulheres.

A literatura recomenda e descreve valores de referência para a verificação da temperatura corporal em quatro locais, a saber: axilar, oral, retal e timpânica. Dentre estas, escolheu-se nesta pesquisa o registro axilar para se obterem os valores de referência de temperatura, uma vez que esta é a que sofre maior influência do ambiente, semelhante ao local que se usou nestes ensaios, que é a região do punho. Conforme Craven e Hirnle (2006), é considerada temperatura normal entre 35.8° e 37.0°C.

3.3.2.2 Respostas Comportamentais Visuais

As respostas comportamentais apresentada pelo telespectador durante a visualização de um vídeo, são consideradas fortes indicadores da cognição humana (KATTI et al., 2011). Segundo Argyle (2013), o interesse das pessoas pode se apresentar com as seguintes reações: sorrir, aumentar a fixação ocular, reduzir as piscadas e movimentar ombros e/ou cabeça. A fixação ocular tem uma relação direta com a atenção, porque geralmente quando alguém está prestando atenção em algo, fixa o olhar neste. Alguns trabalhos recentes que consideram fixações como indicadores do interesse dos seres humanos são as pesquisas de Wang et al.

(2018) e Fang et al. (2019).

Muitas pesquisas em análise implícita de conteúdo afetivo de vídeo assumem que as expressões faciais dos indivíduos revelam com precisão seus sentimentos internos relacionados aos vídeos assistidos. Assim, esses pesquisadores consideram a expressão reconhecida como sendo um descritor emocional daquele vídeo. Entretanto, comportamentos faciais e sentimentos internos de um indivíduo não necessariamente devem ser os mesmos, sabendo que expressões de emoções variam de indivíduo a indivíduo e de contexto a contexto. As abordagens permitem capturar a face do telespectador a partir de uma câmera (ou webcam), realizar o rastreamento facial para, posteriormente a essa fase, realizar a análise da expressão facial que pode resultar no estado emocional predominante ou nos valores percentuais dos estados emocionais.

3.4 Instrumentos de Coleta de Dados dos Participantes

Para coleta dos dados foram utilizados recursos de quatro programas de computador, os quais são descritos a seguir:

Programa 1: (Arduino) Conforme codificações e esquemas descritos nos Apêndices B.2 e C.2, promoveu-se a captura dos dados fisiológicos, a saber:

- condutância da pele, por meio do sensor de resposta galvânica da pele (GSR)[2];
- temperatura, utilizando-se o sensor de temperatura sem contato MLX90614 [3], para os dois primeiros conjuntos de ensaios; e
- saturação de oxigênio e batimentos cardíacos, com o oxímetro de pulso MAX30100 [4].

Programa 2: Realizou-se a captura das informações fornecidas pelo Arduino, enviada do programa anterior, as quais foram organizadas em arquivos de texto contendo o tempo em que ocorreu a coleta e a informação coletada;

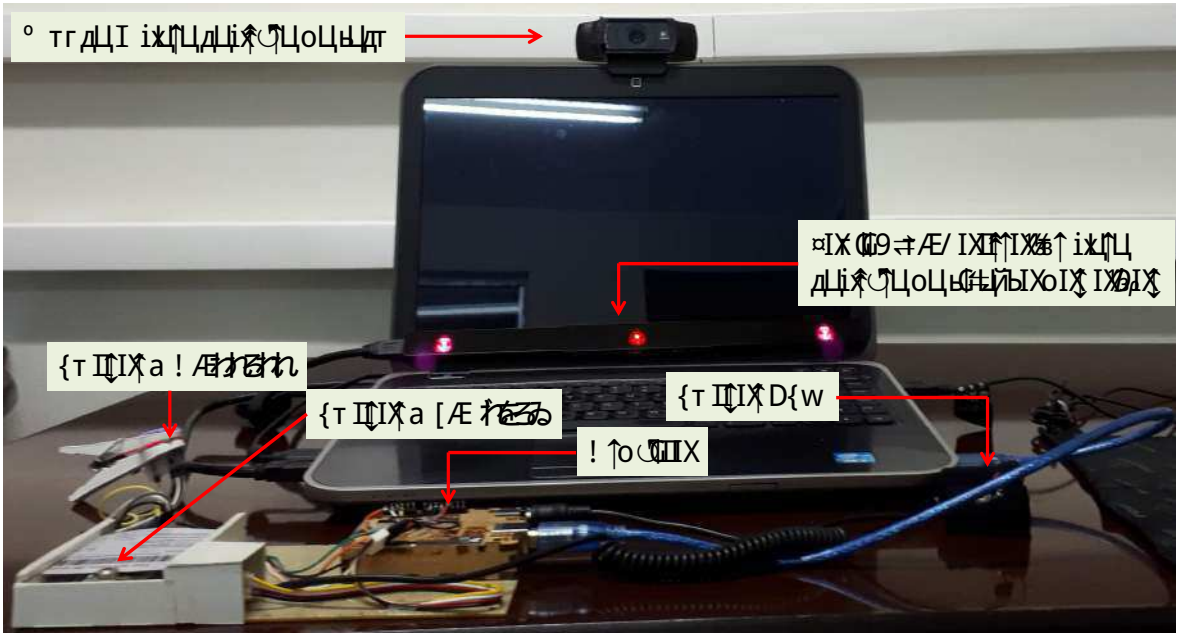
Programa 3: Tratou da captura dos rastros oculares, por intermédio do Tobii EyeX Controller[5] - registros de pontos na tela no qual o participante fixou o olhar. Semelhantemente ao

[2]Manual disponível em: https://www.mouser.com/catalog/specsheets/Seeed_101020052.pdf
[3]Manual disponível em: https://www.sparkfun.com/datasheets/Sensors/Temperature/MLX90614_rev001.pdf
[4]Manual disponível em: https://datasheets.maximintegrated.com/en/ds/MAX30100.pdf
[5]Disponível em: https://help.tobii.com/hc/en-us/articles/209526329-Get-started-with-your-EyeX-Controller

programa 2, compilaram-se os dados em arquivos de texto, tendo sido registrados detalhes como horário de armazenamento e o ponto da tela em que o participante fixou o olhar;

Programa 4: FFmpeg [6](nome criado pela junção da sigla "FF" que significa avanço rápido em português com o grupo de padrões de vídeo MPEG - *Moving Picture Expert Group*) é um programa livremente disponível na Internet que permite converter, gravar e criar fluxo de vídeo e áudio em diversos formatos. A partir deste programa foi gerado um vídeo do participante durante a coleta dos dados. Este é o único programa que não foi desenvolvido no âmbito desta pesquisa. A estrutura física do sistema é mostrada na Figura 3.1

Figura 3.1: Estrutura física desenvolvido para a captura dos dados



Fonte: **Autor**.

Com relação a precisão dos dispositivos utilizados nesta coleta de dados foi apurado que:

- No manual do MLX90614, informa-se que o sensor tem acurácia de 0,01°C, tanto para a temperatura do ambiente, quanto do objeto (participante);
- Os valores da acurácia para os sensores GSR e oxímetro de pulso não foram apresentados em seus respectivos manuais; e
- No site da empresa Tobii[7] afirma-se que os rastreadores oculares geralmente têm uma

[6]Site oficial: https://www.ffmpeg.org
[7]https://www.tobiipro.com/blog/expert-eye-tracking-tips/

precisão em torno de 0,5 a 1 grau do ângulo visual. Na pesquisa de Sarmento, Rangel e Gomes (2016) foi apresentada uma acurácia de 0,5° para o Tobii Eyex Controller encontrado em seu manual de descrições na Web.

3.5 Procedimentos de Coleta dos Dados

A coleta dos dados aconteceu em três conjuntos de ensaios distintos, caracterizados por diferentes coleções de vídeos, bem como diferentes grupos de voluntários. Nestes conjuntos de ensaios, cada participante assistiu à respectiva coleção de vídeos, sendo avaliados segundo suas reações fisiológicas, análise de expressões faciais e rastreamento ocular. Assim, permitindo realizar os aprimoramentos necessários a eficiência da aquisição dos dados na abordagem de sumarização proposta.

3.5.1 Coleta de Dados - Vídeos

Nesta subseção são apresentados os vídeos pertencentes à cada conjunto de ensaios. São contempladas informações relevantes dos vídeos e os comandos utilizados na formatação dos vídeos e na obtenção dos quadros-chave para apresentação aos telespectadores.

3.5.1.1 Seleção e Preparação dos Vídeos para o 1° Conjunto de Ensaios

Seis vídeos da plataforma do YouTube foram utilizados neste conjunto de ensaios. Os vídeos pertenciam a diversos gêneros cinematográficos, a saber: drama, ação, épico e animação. Os excertos de vídeos apresentados aos participantes possuíam resolução de 1920x1080, taxa de 24 quadros por segundo e foram recortados os dez primeiros minutos do vídeo para compor o clipe. Para a execução do recorte, foi utilizada a ferramenta FFmpeg, a partir do seguinte comando:

```
ffmpeg -ss 00:00:00 -t 00:10:00 -i VIDEO.mp4 -acodec copy -vcodec copy CLIPE.mp4
```

Para compor a amostra de vídeos aqui apresentados, foram selecionados vinte vídeos da plataforma do YouTube. De cada clipe obtido, foram extraídos quadros, à taxa de um quadro a cada dois segundos, para serem apresentados ao término da exibição do clipe aos partici-

pantes, totalizando 300 quadros. Para esta finalidade foi utilizada novamente a ferramenta FFmpeg executando o seguinte comando:

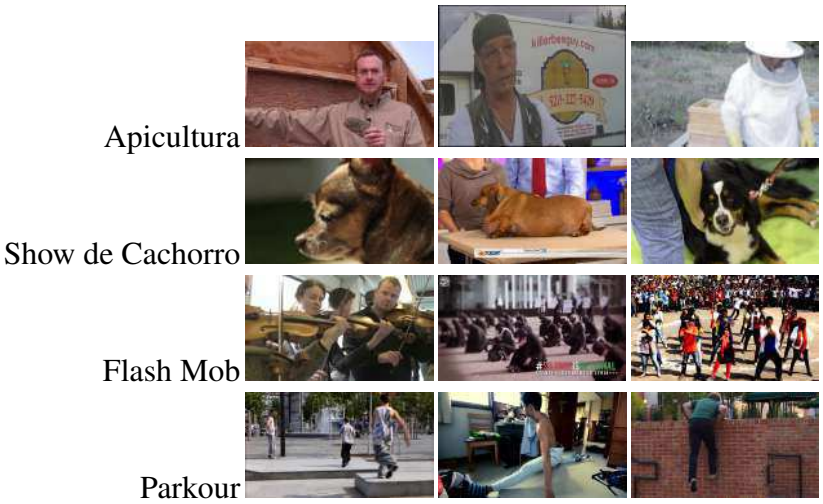
```
ffmpeg -i CLIPE.mp4 -vf fps=1/2 CLIPE%03d.png
```

3.5.1.2 Seleção e Preparação dos Vídeos para o 2° Conjunto de Ensaios

Neste conjunto de ensaios, os vídeos utilizados foram selecionados a partir do banco de dados Title-based Video Summarization - TVSum (SONG et al., 2015), que possui 50 vídeos classificados em 10 categorias, com cinco vídeos cada e resolução máxima de 640x320 cada vídeo. Para participar dos ensaios, foram selecionados doze vídeos do banco de dados, três vídeos em quatro categorias, aqueles com maiores chances de gerarem alterações fisiológicas nos participantes.

Na base, TVSum, os vídeos são identificados conforme a identificação do YouTube, assim permitindo a busca dos vídeos em suas melhores resoluções. Desta forma, todos os vídeos foram apresentados aos participantes a uma taxa de 24 quadros por segundos; destes, a maioria estava na resolução 1280x720. Categorias e instantâneos dos vídeos utilizados nesta coleta são apresentados na Figura 3.2 e informações adicionais relacionadas ao nome original, à identificação utilizada nesta pesquisa, à localização no site, à duração e à categoria são relacionadas no Quadro 3.1.

Figura 3.2: Gêneros e miniaturas dos vídeos coletados do banco de dados TVSum



Fonte: **Autor**.

De forma semelhante ao conjunto de ensaios anterior, dos clipes obtidos foram extraídos

quadros à taxa de um quadro a cada dois segundos para serem apresentados ao término da exibição do clipe aos participantes. Estes variaram de 49 a 118 quadros, dependendo da duração do vídeo. Foram utilizados o mesmo comando e a mesma ferramenta para gerar esses quadros.

3.5.1.3 Seleção e Preparação dos Vídeos para o Terceiro Conjunto de Ensaios

Para compor a amostra de vídeos aqui apresentados, foram selecionados vinte vídeos da plataforma do YouTube. Os critérios utilizados para direcionar as escolhas dos vídeos foram: (i) pertencer a quatro gêneros distintos, (ii) possuir boa resolução (a maioria com 1280x720) e taxa de quadros (24 quadros por segundos); (iii) ter duração máxima inferior a quatro minutos e (iv) apresentar mais do que apenas uma simples tomada. Desta forma, a amostra foi composta de cinco vídeos para cada um dos gêneros, sendo escolhidos dois vídeos de cada gênero para exibição aos participantes, compondo oito vídeos. Comparando-se aos estudos revisados desta área de pesquisa, a quantidade e a soma da duração de todos os vídeos são compatíveis e até superiores, conforme pode ser observado no Quadro 2.2, da página 21.

Quadro 3.1: Descrição dos videoclipes utilizados no Segundo Conjunto de Ensaios

Titulo Original	ID	ID YouTube *	Duração (seg.)	Categoria
Paper Wasp Removal | From the Ground Up	BK1	EE-bNr36nyA	98,132	Apicultura
9/15/11 Killer Bees Kill 1000-lb Hog in Bisbee AZ	BK2	Se3oxnaPsz0	138,867	Apicultura
A Year of Beekeeping	BK3	uGu_10sucQo	167,042	Apicultura
Oliver's Show - Dog's tale	DS1	kLxoNp-UchI	129,997	Show de Cachorro
TODAY- Obie the obese dog works toward weight loss	DS2	jcoYJXDG9sw	199,233	Show de Cachorro
The Dog Show HD 720p	DS3	NyBmCxDoHJU	189,600	Show de Cachorro
CASACL - Flashmob in Copenhagen underground - Peer Gynt	FM1	_xMr-HKMfVA	148,916	Flash Mob
SYRIA SYDNEY FLASH MOB #SILENCEISBETRAYAL	FM2	a-i8LHB1JKU	216,48	Flash Mob
ICC World Twenty 20 Bangladesh 2014, Flash Mob - UITS, Chittagong (Official)	FM3	byxOvuiIJV0	154,53	Flash Mob
Parkour Camp Leipzig	PK1	GsAD1KT1xo8	145,346	Parkour
Singapore Parkour Free Running | JC Boy Late for School	PK2	XkqCExn6_Us	187,888	Parkour
Charlotte Parkour | Charlotte Video Project	PK3	b626MiF1ew4	235,875	Parkour

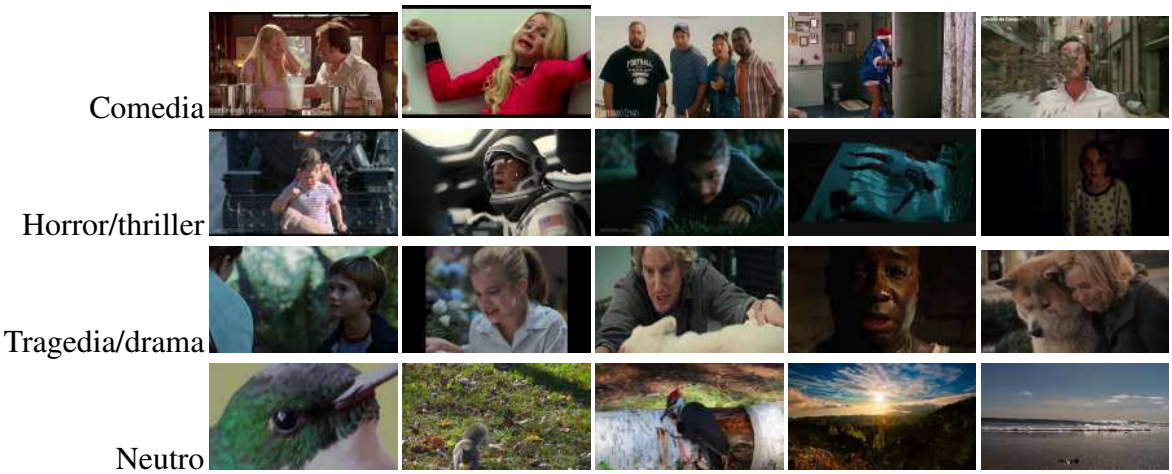
* O endereço completo é formado adicionando o endereço informado ao final de "https://www.youtube.com/watch?v=".
Fonte: **Autor**.

Ação, horror, comédia e drama são gêneros bastante populares. Segundo Xu et al. (2008), as emoções dominantes para filmes de terror e comédia são medo e felicidade, respectivamente. Os mesmos autores consideram difícil designar uma emoção dominante para o drama, mas esse gênero geralmente evoca muitas emoções, enquanto os filmes de ação geralmente atraem a atenção dos telespectadores e sustentam emoções em alta intensidade na maior parte do tempo. No entanto, decidiu-se não usar este último gênero porque pode levar

a emoções muito distintas como medo, raiva e felicidade. Adicionou-se o gênero neutro/relaxante ao conjunto de dados para servir como um contraste e um estado de controle para os outros gêneros.

Na Figura 3.3, são apresentados os gêneros utilizados nesta pesquisa e instantâneos de seus respectivos vídeos. No Quadro 3.2, são descritos os vídeos com suas respectivas informações relacionadas ao nome original, identificação utilizada nessa pesquisa, localização no site, gênero e definição do início da apresentação referente ao vídeo original e duração posterior ao início da apresentação.

Figura 3.3: Gêneros e miniaturas dos vídeos coletados do YouTube



Fonte: **Autor**.

Referente à preparação dos vídeos, foram realizados os seus recortes conforme informações apresentadas no Quadro 3.2 para criação do clipe a ser exibido aos participantes. A ferramenta FFmpeg foi utilizada mais uma vez, empregando-se o seguinte comando:

```
ffmpeg -ss INÍCIO -t DURAÇÃO -i VIDEO.mp4 -acodec copy -vcodec copy CLIPE.mp4
```

De cada clipe obtido, foram extraídos quadros que foram apresentados ao término da exibição do clipe aos participantes para representar cada tomada do vídeo. Novamente, a ferramenta FFmpeg realizou esta tarefa, a partir do seguinte comando:

```
ffmpeg -i CLIPE.mp4 -f image2 -vf “select=’eq(pict_type\,I)’” -vsync vfr CLIPE%03d.png
```

Quadro 3.2: Descrição dos videoclipes utilizados no Terceiro Conjunto de Ensaios

Titulo Original	ID	ID YouTube *	Gênero	Início	Duração
O Amor É Cego (2001) - Namorando Rosemary (4/5) Filme/Clip	AC1	ep6ZtCmsK7I	Comédia	00:00:00	00:02:12
Marcus nas compras (Pt. 2) - As Branquelas	BR1	fw8gKiaXi7c	Comédia	00:00:00	00:01:42
Gente grande (2010) - Tem mais Uma? (6/10) | Filme/Clip HD	GG1	l3SKZK_m9n4	Comédia	00:00:00	00:01:24
Gigolô europeu por acidente(cena do gato safado)	GI1	xN_9GSBIDf8	Comédia	00:00:00	00:02:02
Bruce ganha Seus poderes (HD) (DUBLADO) Todo Poderoso 2003	TP1	ii7FL7t7I98	Comédia	00:00:34	00:02:29
Train! - Stand by Me (2/8) Movie CLIP (1986) HD	CO1	gozRrRCtj6E	Horror/thriller	00:00:00	00:01:58
Interestelar cena da onda gigante	IE1	0mkMFHkhcTo	Horror/thriller	00:00:24	00:02:25
Disturbia (7/9) Movie CLIP - Living in Peace (2007) HD	PA1	toAOUXtlXXc	Horror/thriller	00:00:00	00:02:13
A HORA DO PESADELO - Melhor cena de terror	PE1	0wM1nNx7t4A	Horror/thriller	00:00:06	00:01:07
The Visit	VI1	E6JnsVZ2MRE	Horror/thriller	00:00:00	00:01:05
Mabel Cezar dublando em “A. I. - Inteligência Artificial”	AI1	brUjCnSv0no	Tragédia/drama	00:00:06	00:02:28
Cena Funeral - Meu Primeiro Amor ? Dublado	LO1	xXLoEaCfDV0	Tragédia/drama	00:00:00	00:02:37
Fandub Marley & Eu - Você é um grande garoto! (Cena final)	MA1	c3B18awVWcY	Tragédia/drama	00:00:03	00:01:15
“Acho que posso entender” (À Espera de um Milagre)	MI1	yztusel4vFw	Tragédia/drama	00:01:20	00:02:19
Sempre Ao Seu Lado Dublado Filme completo em Hd	SL1	teJzjgXIP3k	Tragédia/drama	01:20:24	00:02:36
Passarinhos Coloridos Diversos Tipos Lindos Magia Inspiradora da Alma	PC1	CcLd0KtmUyI	Neutro	00:00:00	00:01:22
Paz e harmonia Parques cheios de esquilos passarinhos pombos aves	PH1	Bo-lvEodZfA	Neutro	00:00:00	00:02:38
Pica Pau fazendo ninho em árvores - Música Instrumental ao suave som de Piano	PP1	wW8SciPlI0s	Neutro	00:00:00	00:01:49
O video mais lindo que ja vi...	VL1	aJT9F2oHrSg	Neutro	00:00:00	00:03:08
Vídeo para meditar e relaxar (praia, ondas, mar), com piano relaxante no fundo - Natureza em HD	VM1	W3Ufvm7MqcM	Neutro	00:00:00	00:02:30

* O endereço completo é formado adicionando o endereço informado ao final de “https://www.youtube.com/watch?v=”.
Fonte: **Autor**.

3.5.2 Coleta de Dados - Participantes

Nesta subseção são apresentados os protocolos utilizados no processo da coleta de dados dos participantes pertencentes à cada conjunto de ensaios. São comentados os sensores e dispositivos utilizados e a montagem de cada ensaio durante a coleta de dados dos participantes.

3.5.2.1 Sistematização do Processo de Adequação entre os Participantes e as Formas de Aquisição dos Dados para o 1° e 2° Conjuntos de Ensaios

A coleta de dados dos participantes seguiu o seguinte protocolo:

- Primeiramente, foram obtidos os dados da resposta fisiológica, faciais e de atenção visual do participante, monitorado durante a exibição dos vídeos com o auxílio de webcam, rastreador ocular e dos seguintes sensores: sensor de resposta galvânica da pele, termômetro infravermelho e oxímetro de pulso; e

  - Cada participante foi acomodado confortavelmente sentado a uma distância máxima de 80 cm da tela do computador e do rastreador ocular;
  - O rastreador ocular foi calibrado para cada participante, individualmente, no início de cada seção;
  - Os dedos indicador e médio da mão direita estavam em contato com o sensor de condutância da pele, o dedo indicador da mão esquerda em contato com o oxímetro de pulso e o termômetro infravermelho em contato com a região do pulso do braço esquerdo, enquanto a webcam foi configurada para que o rosto do participante estivesse totalmente enquadrado;
  - Cada participante foi instruído a se posicionar frontalmente à webcam e não realizar movimentos horizontais e verticais da cabeça, bem como não movimentar as mãos, por estarem em contato com os sensores; e
  - Buscou-se um ambiente no qual os ensaios fossem conduzidos, livres de distrações para o participante, assim como, tivesse uma condição de iluminação adequada.

- Posteriormente à exibição de cada clipe, o participante foi convidado a selecionar um conjunto de quadros, para cada vídeo exibido, que considerasse mais relevante para compor um sumário, conforme o roteiro de atividades (APÊNDICE E).
  - Os sensores foram removidos para que o participante realizasse as seleções com as próprias mãos, assim este não necessitou do auxílio do pesquisador;
  - O pesquisador solicitou que cada participante selecionasse quadros do clipe que possibilitassem contar a história que acabou de assistir; e
  - A quantidade de quadros que o participante necessitava selecionar foi estipulada *a priori*, relacionada à quantidade total de quadros. No primeiro conjunto de ensaios, optou-se por solicitar 20% da quantidade total de quadros, enquanto para o segundo conjunto de ensaios o percentual foi de aproximadamente 30%.

3.5.2.2 Sistematização do Processo de Adequação entre os Participantes e as Formas de Aquisição dos Dados para o 3° Conjunto de Ensaios

A coleta de dados dos participantes neste conjunto de ensaios apresentou semelhanças e diferenças em relação aos conjuntos de ensaios anteriores. Por questão de clareza, optou-se por apresentar na íntegra o seguinte protocolo:

- Inicialmente, foram adquiridas as informações da resposta fisiológica, facial e de atenção visual do participante monitorado, enquanto assistia ao vídeo, com o auxílio de webcam, rastreador ocular e dos seguintes sensores: sensor de resposta galvânica da pele e oxímetro de pulso (o termômetro infravermelho foi removido);
  - Cada participante foi posicionado confortavelmente sentado a uma distância aproximada de 80 cm da tela do computador e do rastreador ocular;
  - No início de cada seção, foi realizada a calibragem do rastreador ocular para cada participante, individualmente;
  - Nos dedos indicador e médio da mão direita, foi acoplado o sensor de condutância da pele e no dedo indicador da mão esquerda ficou em contato com o oxímetro de pulso, enquanto que a webcam foi ajustada para que a face do participante estivesse totalmente enquadrada;

  - Cada participante foi informado a se colocar em frente à webcam e não movimentar horizontalmente e verticalmente a cabeça e ainda não movimentar as mãos, por estarem acopladas aos sensores; e
  - Os ensaios foram realizados em ambiente livre de distrações para o participante, assim como também foi consultada a opinião sobre o conforto ambiental (térmica, acústica e iluminação), de modo que ajustes necessários pudessem ser realizados antecipadamente.

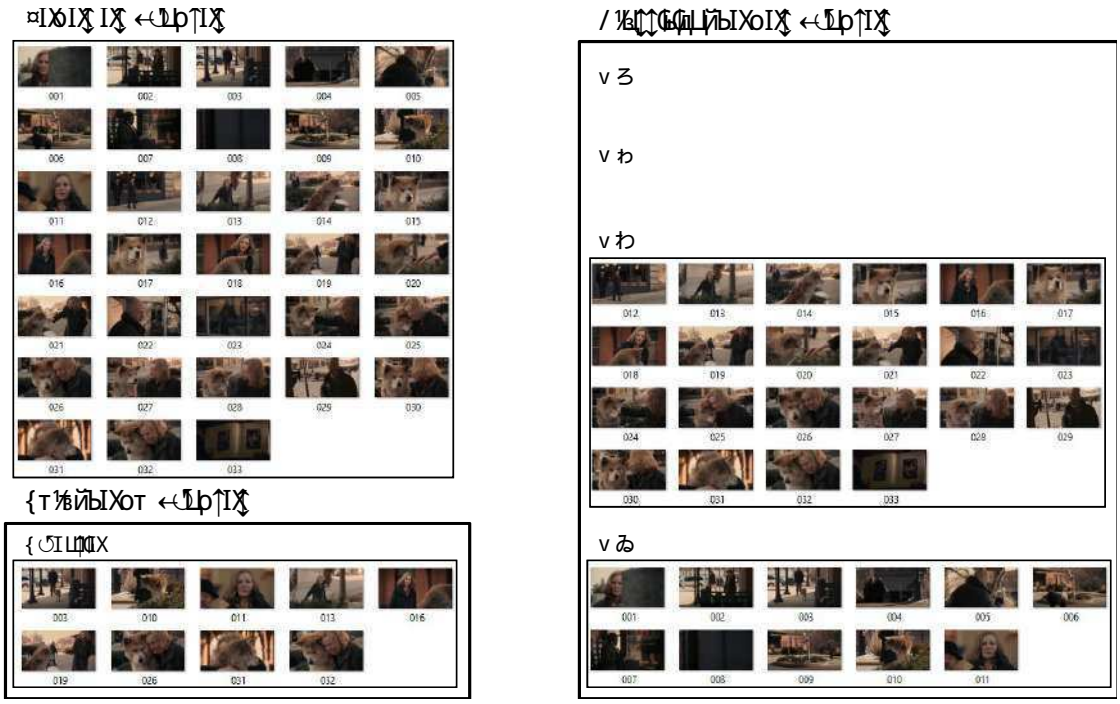
- Anteriormente à exibição do clipe, foi realizada a apresentação de um vídeo com conteúdo neutro para auxiliar a eliminação dos estados emocionais anteriores e ao término desta apresentação, exibiu-se uma sinopse apresentando ao participante, eventos acontecidos no vídeo original até o início do clipe, para os casos que o pesquisador considerou necessária melhor compreensão do clipe; e

- Após a exibição de cada clipe, foi solicitado ao participante que selecionasse um conjunto de quadros que considerasse mais relevante, com o objetivo de compor um sumário e, em seguida, que classificasse cada quadro do conjunto de quadros conforme o Modelo Circumplexo de Emoção de Russell (1980), seguindo a figura que se encontra em Anexo G.1. O roteiro de atividades deste conjunto de ensaios encontra-se no APÊNDICE F

  - Os sensores foram removidos do participante, principalmente para facilitar a seleção e classificação dos quadros;
  - O pesquisador solicitou que o participante selecionasse os quadros do clipe que achasse mais relevantes;
  - A quantidade de quadros que o participante necessitava selecionar também foi estipulada *a priori*, relacionada à quantidade total de quadros. Porém, foi determinado um número mínimo e máximo (aproximadamente 15% e 30% da quantidade total de quadros) de quadros que poderiam ser selecionados; e
  - O participante foi orientado a externar o sentimento ou emoção sentida durante a apresentação de cada tomada representada pelos quadros na atividade de classificação dos quadros.

Na Figura 3.4, apresentam-se os dois processos solicitados ao participante: (i) a seleção de quadros, em que o participante escolheu quais quadros, dentre todos aqueles pertencentes ao vídeo, que deveriam pertencer ao sumário; (ii) a classificação dos quadros, em que o participante classificou cada quadro conforme o quadrante mais relacionado à sua emoção.

Figura 3.4: Processos realizados após a exibição dos vídeos pelo participante



Fonte: **Autor**.

3.5.2.3 Configuração do Sistema para Aquisição dos Dados

A aquisição dos dados, por questões operacionais, teve início antes da exibição do vídeo, com a captura dos registros de rastros oculares e a obtenção dos dados fisiológicos de cada participante. Neste ponto do experimento, obtinha-se a média das 500 primeiras leituras de GSR. A característica GSR foi decomposta em GSR2 (diferença entre o GSR atual e o GSR médio) e GSR1 (quando GSR2 for menor ou igual a 60, o GSR1 é igual a 0 e em caso contrário é igual a 1), a primeira, presente em todos os conjuntos de ensaios e a última, presente nos dois primeiros.

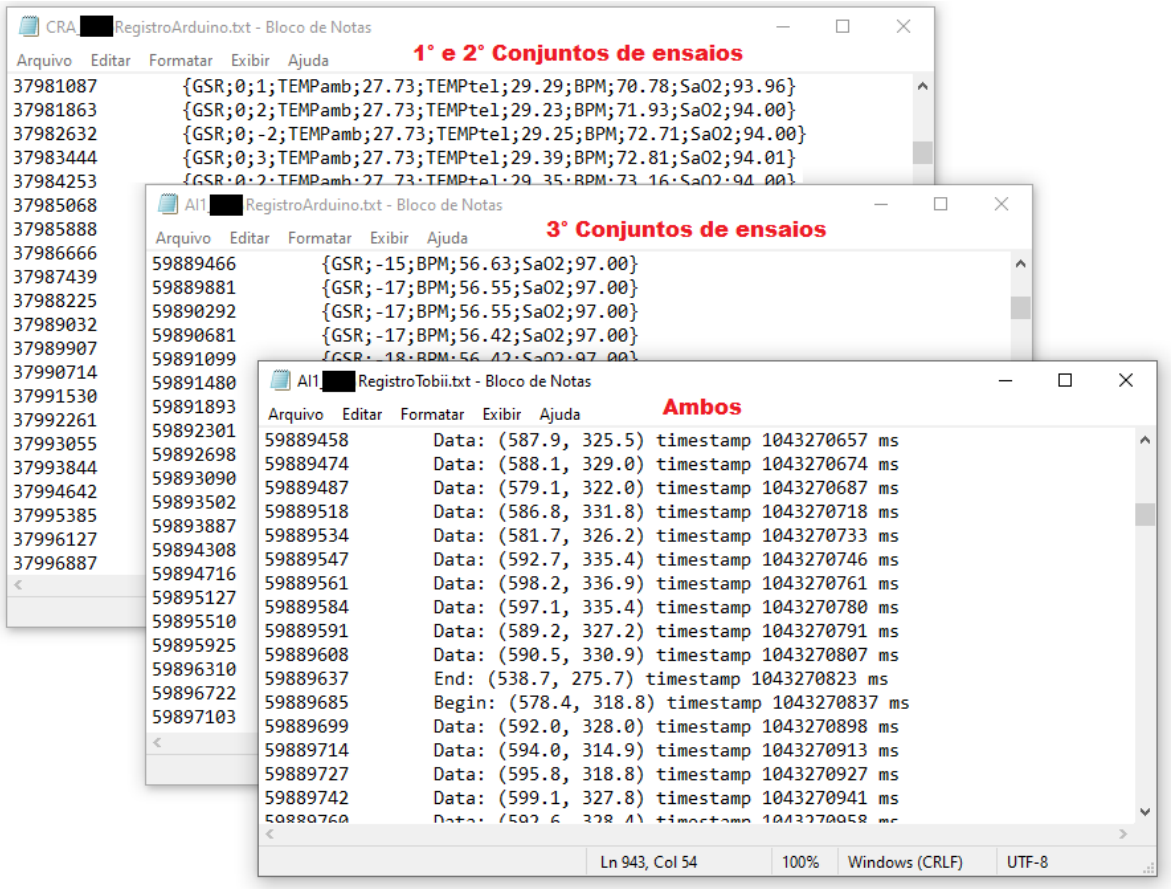
O início e o término da exibição dos clipes delimitaram o período de coleta das imagens faciais do participante pela webcam. Porém, os registros dos rastros oculares e aferição dos dados fisiológicos ainda permaneceram sendo coletados instantes após a finalização da

exibição dos clipes.

A operação de criação dos dados coletados do usuário ocorria ao término da exibição do videoclipe. Cada um dos processos relatados anteriormente foi documentado em registros separados, sendo dois registros textuais e um registro visual (vídeo do usuário). Para evitar o problema de sincronização nos registros textuais, foi adotada a solução de que todos seus registros seriam criados contendo o horário em que foram capturados, segundo o relógio do computador. A solução adotada para o problema anterior, no caso da exibição do videoclipe e da captura do vídeo do usuário, foi a criação de registros textuais contendo o início da exibição e o início da captura obtidos do relógio do computador, respectivamente.

Na Figura 3.5, apresentam-se os dois registros textuais gerados durante a coleta de dados. Os dados contidos nos registros são o horário em que foi capturado e o dado obtidos pelo(s) sensor(es) naquele horário.

Figura 3.5: Registros textuais gerados na coleta de dados



Fonte: **Autor**.

Para cumprir as atividades desta fase, foi criado um *script* que executava algumas ope-

rações necessárias para que o sistema funcionasse, bem como realizava as chamadas aos programas que fizeram essas capturas em separado. Os programas utilizados nesta fase foram:

- O programa de captura e criação dos dados de registros de rastreamentos oculares que foi baseado em exemplos obtidos na biblioteca do rastreador;
- O programa responsável pela obtenção e criação dos registros fisiológicos que foi adaptado de um leitor de portas seriais obtido na Internet. O programa gravado no Arduino realizava o processamento necessário e enviava uma *string* contendo as informações obtidas para a porta serial do computador; e
- O programa ffmpeg foi utilizado para a captura do vídeo do usuário utilizando o mesmo número de quadros por segundo e tempo de duração do vídeo assistido.

3.6 Análises dos Dados

Os dados foram analisados em três etapas, quais sejam: sincronização, normalização e utilização dos dados.

3.6.1 Sincronização

A fase iniciava com a criação de um registro de fixações. Os intervalos das fixações foram obtidos dos intervalos entre um "begin" e um "end" do registro de rastros oculares, sendo um registro de fixação composto pelo seu início e fim. Para a obtenção das fixações relacionadas a intervalos de meio, um, um e meio e dois segundos de duração, respectivamente, foram aplicados filtros referentes a cada um dos intervalos de duração, excluindo-se os intervalos retornados do processo que possuíam duração total inferiores a meio, um, um e meio e dois segundos de duração, respectivamente. Assim, determinaram-se valores de fixações para cada uma dessas durações.

Foi realizada a criação de um registro sincronizado dos dados fisiológicos e das fixações utilizando o tempo em que ocorreu a coleta referente à exibição de cada um dos quadros do vídeo. Em seguida, os vídeos gerados do participante foram submetidos ao programa OpenFace, apresentado em Baltrusaitis, Robinson e Morency (2016) para extração de seus AU do

Facial Action Coding System (FACS), que teve como saída um registro textual contendo seus valores para cada quadro do vídeo. Estes dois registros foram fundidos em um e, posteriormente, foi gerado o registro de sincronização geral dos dados referentes a intervalos de dois segundos. Para o 3° conjunto de ensaios, ainda foram adicionadas as informações duração e o número de fixações ao registro de sincronização geral dos dados. Esta fase será mais bem detalhada nos próximos parágrafos.

3.6.1.1 Registro Sincronizado

A Figura 3.6 contém um registro sincronizado em cada conjunto de ensaios. Como pode ser visto, os registros estão organizados por quadro, dessa forma, cada quadro apresentou algum valor para as variáveis analisadas nessa pesquisa.

Os registros textuais dos rastros oculares e registros fisiológicos foram sincronizados utilizando-se os horários registrados em que ocorreu a coleta e os registros textuais contendo o início da exibição e início da captura, conforme descrito na fase de coleta dos dados. Durante a sincronização, o horário de início da exibição do videoclipe foi utilizado como base para a sincronização dos dados. Dessa forma, o horário registrado foi convertido para ser relativo a cada quadro do vídeo.

3.6.1.1.1 Dados Fisiológicos

O esquema utilizado para a sincronização e extração dos dados fisiológicos foi sistematizado na seguinte sequência:

- Os dados anteriores ao horário-base do registro eram ignorados e não foram extraídos para o registro global sincronizado;
- O primeiro dado do registro com horário posterior ao horário-base foi extraído e registrado com o valor do horário base no registro global sincronizado; e
- Os demais dados do registro global sincronizado foram obtidos do dado do registro com horário posterior ao horário obtido da formula $Horario = HorarioBase + \frac{NumFrame-1}{fps}$, na qual $NumFrame$ correspondeu à posição relativa ao quadro em que o vídeo assistido estava sendo processado e fps correspondeu ao número de quadros

por segundo do vídeo assistido. Esse dado foi armazenado no registro sincronizado com o horário obtido da formula descrita anteriormente.

Figura 3.6: Registros globais textuais dos conjuntos de ensaios

frame	GSR1	GSR2	TEMPamb	TEMPtel	BPM	SaO2	Fix0_5s	Fix1s	Fix1_5s	Fix2s
1	0	0	30.27	33.85	67.65	94.36	1	1	1	1
2	0	0	30.27	33.85	67.65	94.36	1	1	1	1
3	0	0	30.31	33.85	73.07	94.39	1	1	1	1
4	0	0	30.31							
5	0	0	30.31							
6	0	0	30.31							
7	0	0	30.31							
8	0	0	30.31							
9	0	0	30.31							
10	0	0	30.31							
11	0	0	30.31							
12	0	0	30.31							
13	0	0	30.31							
14	0	0	30.31							
15	0	0	30.31							
16	0	0	30.31							
17	0	0	30.31							
18	0	0	30.31							

1° e 2° Conjuntos de Ensaios

frame	GSR2	BPM	Fix0_5s	Fix1s	Fix1_5s	Fix2s
1	-13.500000	56.26000	0.0000000	0.00000000	0.00000000	0
2	-9.298507	57.13993	0.4477612	0.20149254	0.00000000	0
3	-16.616000	56.35800	0.3880000	0.11200000	0.00000000	0
4	-23.714286	55.69000	0.3571429	0.00000000	0.00000000	0
5	-30.180000	55.40568	0.1840000	0.13200000	0.00000000	0
6	-41.312000	55.08356	0.3560000	0.11200000	0.00000000	0
7	-47.801282	54.55885	0.1730769	0.00000000	0.00000000	0
8	-51.307692	54.20538	0.9487179	0.61538462	0.00000000	0
9	-53.340000	53.64153	0.1200000	0.03333333	0.00000000	0
10	-35.879781	54.77448	0.5519126	0.30054645	0.00000000	0
11	-38.000000	55.86390	1.0000000	0.87804878	0.87804878	0
12	-39.293578	55.40835	0.6513761	0.61467890	0.06422018	0
13	-42.675676	55.14304	0.7905405	0.61486486	0.27027027	0

3° Conjunto de Ensaios

Fonte: **Autor**.

3.6.1.2 Registro de Sincronização Geral dos Dados

Um registro sincronizado geral contendo os valores do registro sincronizado e os valores das características AU, obtidos da extração das AU da face do telespectador foi gerado. Posteriormente, os valores médios referente ao intervalo (dois segundos, para os casos do 1° e 2° conjuntos de ensaios e da duração das respectivas tomadas, no caso do 3° conjunto de ensaios) desse registro sincronizado geral foram graduados para a obtenção do Registro de Sincronização Geral dos Dados.

3.6.1.2.1 Extração e Registro das Características AU da Face do Telespectador

O registro visual obtido na etapa anterior precisou ser convertido em um registro textual relacionado ao estado afetivo do telespectador. Assim, essa fase foi caracterizada pelo processo

de extração das características AU da face do telespectador. O extrator de AU processou o vídeo do usuário durante a visualização do vídeo, rastreando e realizando o processamento necessário da face do usuário quadro-a-quadro desse vídeo e retornando um registro texto contendo a informação da intensidade para cada AU, nos respectivos quadros do vídeo.

Os vídeos contendo as faces dos participantes foram processados por meio do kit de ferramentas OpenFace 2.0 (BALTRUSAITIS et al., 2018). O OpenFace 2.0 está disponível para *download* no repositório github[8]. Este *kit* de ferramentas suporta experimentos de análise de comportamento facial. A partir do rico registro obtido, que inclui, dentre outros, a estimativa de cabeça e olhos, e o reconhecimento de um subconjunto de AU pelo OpenFace 2.0. Nesta pesquisa foram consideradas apenas as características relacionadas à intensidade de AU gravadas quadro-a-quadro no vídeo.

O OpenFace executa a detecção e o alinhamento de faces (quadro-a-quadro) usando um SVM (*Support Vector Machine*), seguido pelo CE-CLM (*Convolutional Experts Constrained Local Model*). Correções relacionadas com a rotação do plano da face foram executadas. As características de aparência são, então, extraídas, usando Histogramas de gradientes orientados (HOGs) (FELZENSZWALB et al., 2010).

O *kit* de ferramentas OpenFace fornece a estimativa de intensidade de AU, por meio de *Support Vector Regression* (SVR), e detecção de ocorrências de AU, obtidas mediante SVM. Em ambos os casos, foram utilizados núcleos lineares. Como a ocorrência de AU é naturalmente desbalanceada, uma subamostragem de amostras AU negativas dos dados de treinamento foi realizada, com o objetivo de equilibrar o número de amostras positivas e negativas nos dados de treinamento (BALTRUSAITIS; MAHMOUD; ROBINSON, 2015; BALTRUSAITIS et al., 2018).

O subconjunto de AU reconhecido pelo sistema é: 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26 e 45 de acordo FACS. O FACS é um sistema baseado no conhecimento anatômico dos músculos faciais e suas configurações de movimentos para medir o comportamento facial inicialmente proposto por Ekman e Friesen em seu estudo intitulado *"Universal and cultural differences in facial expression of emotion"* (HILDEBRANDT; OLDERBAK; WILHELM, 2015). FACS refere-se aos pequenos movimentos da face que resultam de emoções e outros estados fisiológicos.

[8]https://github.com/TadasBaltrusaitis/OpenFace

Os resultados apresentados no artigo Baltrusaitis et al. (2018) afirma que a abordagem SVR-HOG empregada no OpenFace 2.0 superou as abordagens mais complexas e recentes na detecção de AU no conjunto de dados DISFA (MAVADATI et al., 2013), com um coeficiente de correlação Pearson de 0,59.

Três tipos de registro sincronizado foram utilizados nesta pesquisa: o registro sincronizado dos dados fisiológicos e características fixações, o registro sincronizado quadro-a-quadro e o registro sincronizado geral dos dados, conforme explicado a seguir. O registro sincronizado quadro-a-quadro foi criado do subconjunto de AU relatado anteriormente, que já estava organizado quadro-a-quadro, agrupado com o registro sincronizado dos dados fisiológicos e características fixações.

O registro sincronizado geral dos dados foi obtido do valor médio de cada característica do registro sincronizado quadro-a-quadro, referente ao intervalo utilizado no conjunto de ensaios (dois segundos, para os casos do 1° e 2° conjuntos de ensaios, e da duração das respectivas tomadas, no caso do 3° conjunto de ensaios), e para o 3° conjunto de ensaios acrescentado às informações rótulo emocional do quadro, duração e o número de fixações.

3.6.2 Normalização

Para cada participante, os registros de todos os vídeos visualizados foram primeiramente agrupados obtendo-se um vetor de características $\Phi = \{F_1, F_2, \ldots, F_i\}$, em que F_i é o conjunto de características do vídeo i e representado como:

$$F_i = \{f(i,1), f(i,2), \ldots, f(i,\lambda)\} \text{ e} \tag{3.1}$$

$$f(i,j) = \{f^1_{(i,j)}, f^2_{(i,j)}, \ldots, f^\eta_{(i,j)}\} \tag{3.2}$$

em que λ é o número de quadros-chave pertencentes ao vídeo i, $f(i,j)$ é o conjunto de características pertencentes ao quadro-chave j do vídeo i, $f^l_{(i,j)}$ é o valor da característica l para o quadro-chave j do vídeo i e η é o número de características usadas nesta pesquisa, igual a 26 (vinte e seis), das quais duas foram características fisiológicas, dezessete foram AU, seis foram fixações e um rótulo emocional do quadro.

Para cada característica do vetor Φ, foi aplicada a normalização *Zero-mean*, ver Equação

3.3.

$$\check{f}^{l}_{(i,j)} = \frac{f^{l}_{(i,j)} - \bar{f}^{l}}{\sigma}, \tag{3.3}$$

em que $\check{f}^{l}_{(i,j)}$ é o valor da característica normalizada l, $\bar{f}^{l}$ é a característica média l, e σ é o desvio padrão da característica l. Assim, também criou-se o valor para o vetor de características normalizadas ($\check{\Phi}$), conforme Equações 3.7 a 3.6.

$$\check{f}(i,j) = \{\check{f}^{1}_{(i,j)}, \check{f}^{2}_{(i,j)}, \check{f}^{3}_{(i,j)}, ..., \check{f}^{\eta}_{(i,j)}\} \quad , \tag{3.4}$$

$$\check{F}_i = \{\check{f}(i,1), \check{f}(i,2), \ldots, \check{f}(i,\lambda)\} \quad \text{e} \tag{3.5}$$

$$\check{\Phi} = \{\check{F}_1, \check{F}_2, \ldots, \ldots, \check{F}_i\} \quad , \tag{3.6}$$

em que $\check{f}(i,j)$ é o conjunto de características normalizadas pertencentes ao quadro-chave j do vídeo i, e $\check{F}_i$ é o conjunto de características normalizadas do vídeo i.

3.6.3 Utilização

Nesta subseção são apresentadas as ferramentas utilizadas para obtenção dos sumarizadores apresentados nesta pesquisa. São realizados breves comentários sobre os modelos de máquina de aprendizagem, estratégias de reamostragem e seleção de características utilizadas nos sumarizadores. Nos experimentos, os modelos de máquina de aprendizagem, estratégias de reamostragem e o seletor de características Boruta estão todos implementados como pacotes (**caret** e **Boruta**) do ambiente estatístico R.

3.6.3.1 Modelos de Máquina de Aprendizagem

A aprendizagem de máquina pode ser vista como um ramo da Ciência da Computação que fornece aos computadores a capacidade de "aprender" sem serem explicitamente programados (SAMUEL, 1959; ALPAYDIN, 2009). Na literatura, redes Bayesianas (JOHO et al., 2009), KNN (DAMMAK; WALI; ALIMI, 2015), SVM e Redes Neurais são exemplos de técnicas empregadas nesta área.

Para o treinamento, concentrou-se em aperfeiçoar a precisão da classificação. Estimou-

se o desempenho de um dado modelo utilizando validação cruzada com 10 dobras e 3 replicações (*repeated n-fold cross validation, with 10 folds and 3 replicates*). Os ajustes dos parâmetros foram tratados automaticamente pelo pacote **caret** do ambiente estatístico de código aberto R[9]). O **caret** (abreviação de *Classification And REgression Training*) tem várias funções focadas em simplificar o processo de construção e avaliação de modelos complexos de classificação e regressão, bem como a seleção de características e outras técnicas (KUHN, 2008).

3.6.3.2 Estratégias de Reamostragem

Um conjunto de dados desbalanceados é caracterizado por possuir mais instâncias pertencentes a uma classe do que a outra. Normalmente é considerada uma classe positiva ou minoritária e uma classe negativa ou majoritária. Este tipo de conjunto de dados cria problemas para vários campos, como mineração de dados, porque os algoritmos padrão de aprendizado de máquina não lidam bem com dados desbalanceados e tendem a realizar suas classificações em direção à classe negativa/majoritária (ROUT; MISHRA; MALLICK, 2017; EL-AMIR; EL-FIQI, 2019).

Segundo Aurelio et al. (2019), El-Amir e El-Fiqi (2019), existem vários métodos propostos para lidar com conjuntos de dados desbalanceados e podem ser divididos em três grupos: nível de dados, nível de algoritmo e abordagens sensíveis ao custo. Na primeira, abordagem no nível de dados ou técnicas de reamostragem, os dados utilizados para o treinamento são reamostrados para melhorar o balanceamento das duas classes. No nível de algoritmo, os algoritmos de aprendizado são ajustados para tomar uma decisão de forma tendenciosa para a classe minoritária. As técnicas sensíveis ao custo mesclam as técnicas de reamostragem e de algoritmo, reduzindo o custo de maneira abrangente e aumentando o custo nas classificações incorretas a instâncias da classe minoritária.

Sabendo-se que o número de quadros pertencentes ao sumário é muito menor do que aqueles que não devem pertencer ao sumário, este é um conjunto de dados desbalanceado, desta forma, um conjunto de dados que afeta negativamente o desempenho das máquinas de aprendizagem e, portanto, deve ser tratado. O tratamento utilizado foi com estratégias de reamostragem. As seguintes estratégias de reamostragem foram usadas: *Under*, *SMOTE*

[9]https://www.r-project.org/

e *ROSE*. Para as avaliações experimentais, o conjunto de quadros-chave selecionados (que devem pertencer a um resumo) é a saída desejada.

Conforme Chawla et al. (2002), a estratégia *Under* executa a subamostra da classe majoritária criando um subconjunto aleatório dos dados dessa classe para corresponder à população da classe minoritária e a população dessa classe é mantida. Para o mesmo autor, o *SMOTE (Synthetic Minority Oversampling TEchinque)* cria um conjunto de dados no qual a classe minoritária é aumentada, obtendo dados "sintéticos", interpolando valores entre uma instância e vizinhos mais próximos selecionados aleatoriamente e reduzindo a classe majoritária. A estratégia *ROSE (Random Over-Sampling Examples)*, de acordo com Lunardon, Menardi e Torelli (2014), implementa uma estratégia similar a *SMOTE*, mas os dados "sintéticos" são distribuídos uniformemente na vizinhança da classe minoritária.

3.6.3.3 Seleção de Características

A operação de seleção de características reduz o número de características no conjunto de dados, obtendo um conjunto ideal de características que maximize um determinado objetivo e retornando o conjunto de dados modificado (POST; PUTTEN; RIJN, 2016; OLSON; MOORE, 2019). Conforme apresenta Post, Putten e Rijn (2016), melhor interpretação, generalização e velocidade de aprendizado são alguns dos propósitos da seleção de características.

A relevância das características relacionadas aos sumários obtidos dos telespectadores foi determinada usando o método **Boruta**. O Boruta é um método de seleção de características baseado no algoritmo de aprendizagem *Random Forest*, capaz de determinar a relevância das características de maneira imparcial e estável (KURSA; JANKOWSKI; RUDNICKI, 2010). A essência do algoritmo consiste basicamente em realizar uma cópia aleatória dos dados a qual é mesclada com o original e o classificador é construído para esses dados estendidos. A importância da variável nos dados originais é avaliada quando comparada com aquela obtida pelas variáveis aleatórias. Variáveis que ganharam importância sobre as variáveis aleatórias são consideradas mais relevantes (KURSA; RUDNICKI, 2010).

3.7 Estruturação das Avaliações Experimentais

Na avaliação experimental da Seção 4.2, (Investigação da Sumarização Automática de Vídeos usando Características Fisiológicas, Faciais e da Atenção Visual dos Telespectadores), inicialmente eram removidos todos os registros referentes aos vídeos de conteúdo neutro, posterior a esta fase, foi realizada uma redução de características no vetor segundo os tipos de características obtidas dos telespectadores, a saber: características fisiológicas, faciais e de atenção visual. Posteriormente, foram construídos vinte conjuntos de treinamento e teste, utilizando-se 60% dos registros das características para treinamento. Os demais dados foram utilizados no teste. Para o treinamento, foram utilizadas as máquinas de aprendizagem LogitBoost, pls, simpls, *naive_bayes*, *Random Forest* e *Support Vector Machines (SVM)* e as estratégias de reamostragem *Under*, SMOTE e ROSE.

Um seletor aleatório foi construído para servir como base para a comparação, no processo de sumarização. O seletor aleatório apresentou os quadros que considerou relevante e comparou com os quadros informados pelo telespectador como importante. Desta forma, obtinha-se o valor do CUS_A. Este processo foi realizado vinte vezes para cada participante, gerando *a posteriori* o conjunto dos CUS_A do seletor aleatório.

Cada máquina de aprendizagem/estratégia de reamostragem foi utilizada nos vinte conjuntos de treinamento e teste para construir o conjunto dos CUS_A a partir dos modelos aprendidos. Os valores obtidos para CUS_A de cada máquina foram comparados estatisticamente com os valores do CUS_A da seleção aleatória, gerando os resultados desta comparação ao final. Apresenta-se esse processo descrito na Figura 3.7.

Na Figura 3.8, é apresentado um diagrama que representa o processo de avaliação experimental da Seção 4.3 - Análise da Relevância das Características Fisiológicas, Faciais e da Atenção Visual dos Telespectadores para a Sumarização Automática de Vídeos Digitais. De forma semelhante à avaliação anterior, os registros de vídeos de conteúdo neutro foram removidos. Em seguida, foi realizada a construção de vinte conjuntos de treinamento com 60% dos registros das características e 40% para teste. Os conjuntos de treinamento e de teste foram utilizados com o classificador *Random Forest* (LIAW; WIENER, 2002) com a estratégia de reamostragem Under, utilizando-se todas as características.

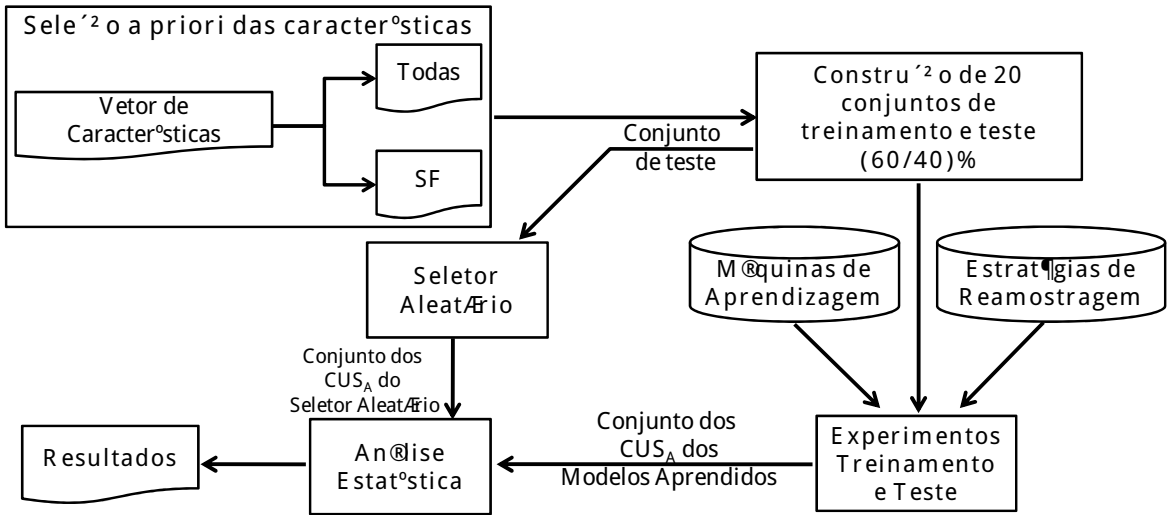
Posteriormente, empregando-se o seletor de características Boruta, foi atribuído um es-

core de relevância para cada característica por experimento (S_k^l), em que k e l são a identificação do experimento e característica, respectivamente. A atribuição do valor a S_k^l foi feita da seguinte forma: se o seletor indicasse "Rejeitado", era atribuído o valor 0; se "Tentativa" fosse igual a 0,5; e atribuia-se para as demais indicações, o valor 1. Subsequentemente, um escore da relevância média para cada característica foi obtido, após 20 experimentos, de acordo com a Equação 3.7.

$$\bar{S}^l = \frac{\sum_{k=1}^{\tau} S_k^l}{\tau} \tag{3.7}$$

em que $\bar{S}^l$ é o escore da relevância média para a característica l e τ é o número de experimentos realizados, que nesta pesquisa foi de 20.

Figura 3.7: Processo de avaliação experimental apresentado na Seção 4.2



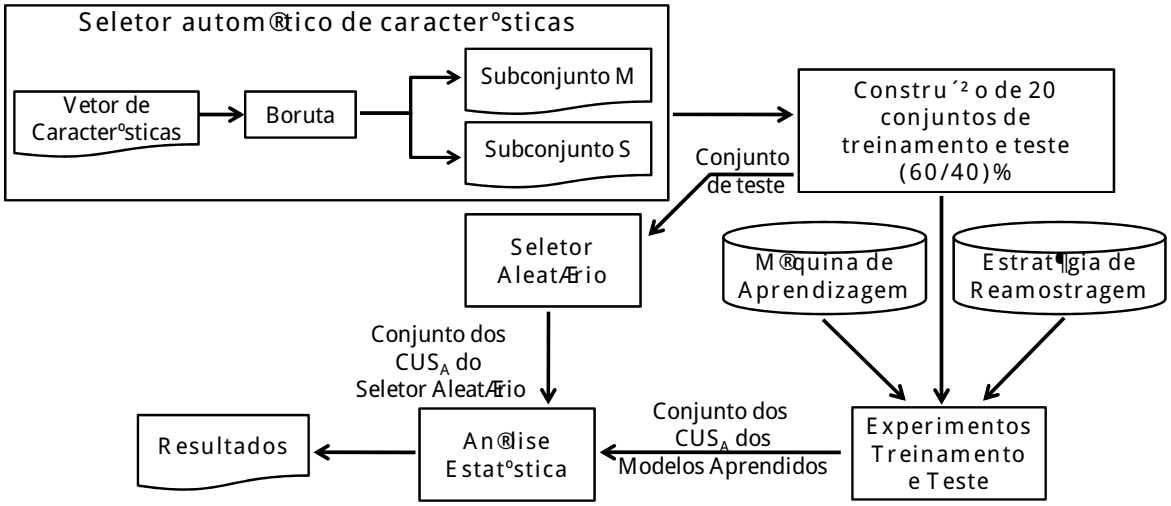
Fonte: **Autor**.

Subsequentemente, calculou-se a relevância média normalizada por participante, que segue a Equação 3.8.

$$\breve{\bar{S}}^l = \frac{\bar{S}^l}{\sum_{k=1}^{\eta} \bar{S}^k} \tag{3.8}$$

em que $\breve{\bar{S}}^l$ é o escore de relevância média normalizada da característica l, $\bar{S}^l$ é o escore de relevância média da característica l e $\sum_{k=1}^{\eta} \bar{S}^k$ é o somatório de todos os escores de relevância média.

Figura 3.8: Processo de avaliação experimental apresentado na Seção 4.3



Fonte: **Autor**.

Assim, obteve-se uma nova escala na qual a soma de $\breve{\breve{S}}^l$ é igual a 1 (100%). Então, $\breve{\breve{S}}^l$ foi classificado de forma decrescente e foi atribuído ao vetor Ψ. O escore de relevância para cada característica foi obtido a partir de Ψ^h, em que h varia de 1 a η. Para a obtenção do vetor de características mais relevantes relacionado a um limiar $theta$ foi utilizada a Equação 3.9.

$$\nu = \arg\max_{h} \sum_{h=1}^{\eta} \Psi^h \leq \theta \tag{3.9}$$

em que ν é o vetor de características mais relevantes relacionado a um limiar θ.

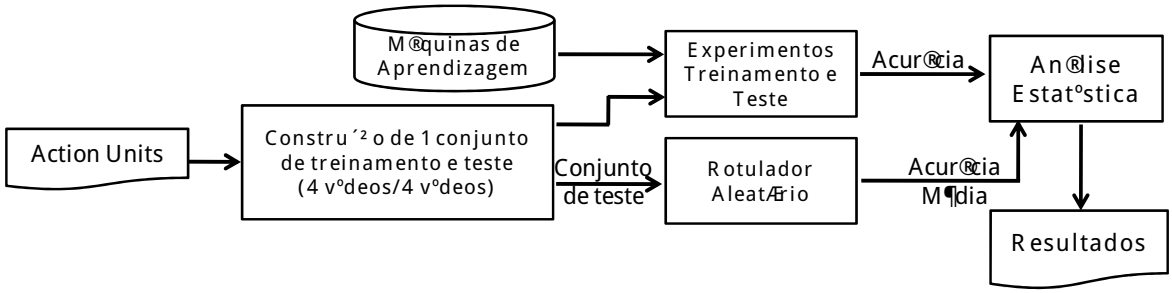
O subconjunto de características M é composto por aquelas que apresentaram um escore de relevância média maior ou igual a um limiar e o subconjunto S é composto pelas características que apresentaram um escore de relevância média normalizada, menor ou igual a outro limiar.

Dando continuidade à estruturação das avaliações experimentais, foram realizados os treinamentos e testes utilizando os 20 conjuntos apresentados anteriormente para cada subconjunto de características M e S analisado. Para o treinamento, foi utilizado o mesmo classificador (*Random Forest*), com a mesma estratégia de reamostragem (*Under*). Os processos de construção do seletor aleatório, análise estatística e geração dos resultados seguiram procedimento idêntico àquele apresentado no processo de avaliação experimental descrito anteriormente.

Na Seção 4.4 (Estudo da Relação entre Expressões Faciais e Estados Emocionais Induzidos pela Apresentação de Conteúdos Multimídia), foi estudada a relação entre as AU e emoções autorreportadas. Neste sentido, AU (ou um subconjunto delas) foram utilizadas como as entradas para modelos de máquinas de aprendizagem e os rótulos emocionais, como as saídas, nestes experimentos. Os resultados dos experimentos objetivaram comparar rotulagens de máquinas de aprendizagem com aquelas de um processo aleatório. Os quadrantes emocionais de referência para estes experimentos foram atribuídos por cada participante aos quadros do vídeo. Para tanto, três experimentos foram propostos.

Na Figura 3.9, é apresentado um diagrama que representa o primeiro experimento. Neste experimento foram selecionados quatro vídeos para treinamento e quatro para teste. Os vídeos utilizados para treinamento foram aqueles que apresentaram a maior quantidade de tomadas rotuladas em cada um dos quadrantes emocionais e os demais vídeos foram usados para formar o conjunto de teste. As máquinas de aprendizagem utilizadas nos treinamentos e testes foram K-Nearest Neighbor, Support Vector Machine, Random Forest, Neural Network e LogitBoost. Após o treinamento e de posse do vetor de características obtiveram-se os rótulos dos quadros.

Figura 3.9: Processo de avaliação experimental apresentado no 1° Experimento da Seção 4.4



Fonte: **Autor**.

A acurácia para cada máquina e participante foi computada a partir da comparação entre os rótulos obtidos do teste e aqueles obtidos do telespectador. De maneira semelhante, foi computada a acurácia para o modelo aleatório, posteriormente à rotulagem aleatória. Este processo foi realizado trinta vezes para cada telespectador, gerando, por conseguinte, a acurácia média do rotulador aleatório. A análise estatística foi realizada comparando os resultados obtidos das acurácias das máquinas com aquelas obtidas do rotulador aleatório.

No segundo e terceiro experimento, os quadros de todos os vídeos foram agrupados e

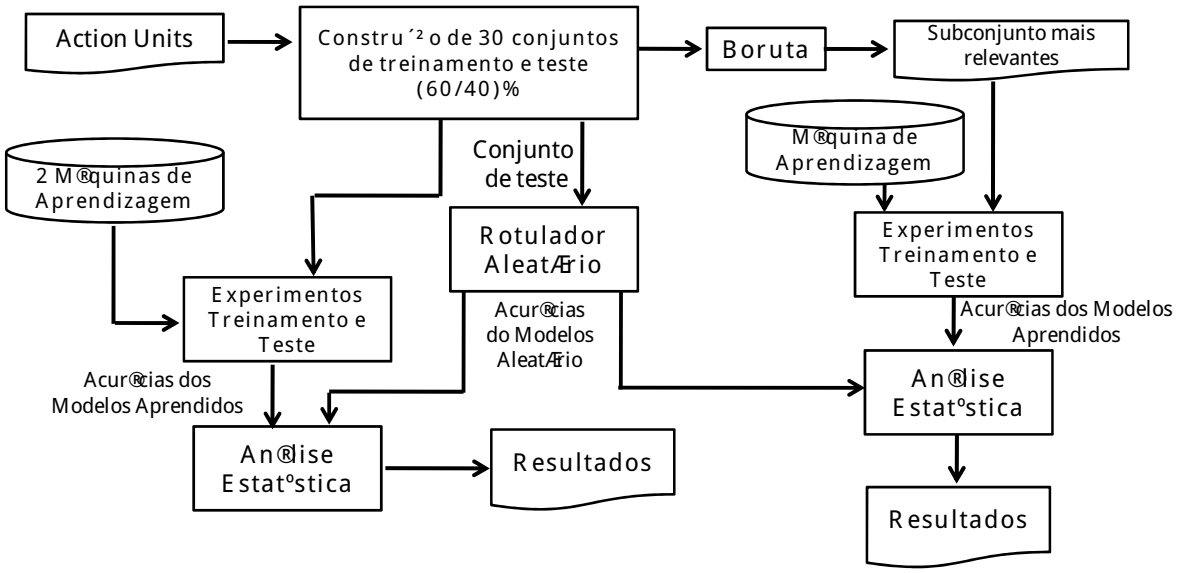
foram construídos 30 conjuntos com 60% dos quadros selecionados aleatoriamente para o treinamento e os quadros remanescentes para teste. A acurácia para 30 rotulagens aleatórias para cada participante foi obtida.

No segundo experimento, as máquinas de aprendizagem utilizadas foram aquelas que apresentaram os melhores resultados no experimento anterior: SVM e Random Forest. Os rótulos dos quadros foram obtidos após treinamento e de posse do vetor de características. Realizou-se a comparação entre os rótulos obtidos do teste e os rótulos do telespectador para obter a acurácia. Este processo foi repetido 30 vezes para cada máquina e participante.

No terceiro experimento foi utilizado o seletor de características Boruta, para reduzir o número de características utilizadas no treinamento pela máquina de aprendizagem Random Forest. As demais etapas são semelhantes àquelas ocorridas no segundo experimento.

Ambas as análises estatísticas foram realizadas comparando os resultados obtidos das acurácias das máquinas com aquelas do rotulador aleatório. Esses processos estão descritos na Figura 3.10.

Figura 3.10: Processos de avaliação experimental apresentado nos 2° e 3° Experimentos da Seção 4.4



Fonte: **Autor**.

Nas etapas de sincronização, normalização e avaliação experimental, foi utilizado o ambiente estatístico *open source* R[10] (LIAW; WIENER, 2002).

[10]https://www.r-project.org/

3.8 Aspectos Éticos

Esta pesquisa foi respaldada nos princípios éticos preconizado pela Resolução nº466/2012 Conselho Nacional de Saúde do Ministério da Saúde (CNS/MS), que norteia as pesquisas envolvendo seres humanos, englobando referenciais da bioética, tais como: justiça, autonomia, equidade, beneficência e não maleficência, assegurando à comunidade acadêmica, aos participantes da pesquisa e ao Estado seus direitos e deveres. Asseguraram-se aos participantes da pesquisa respeito, reconhecimento da sua vulnerabilidade, deixando-os livres para decidir se contribuiriam e permaneceriam ou não na pesquisa, mediante sua manifestação expressa, livre e esclarecida.

A pesquisa foi apreciada eticamente pelo Comitê de Ética em Pesquisas com seres humanos do Hospital Universitário Alcides Carneiro da Universidade Federal de Campina Grande (CEP/HUAC/UFCG), tendo sido aprovada em 24 de abril de 2018, sob o Certificado de Apresentação para Apreciação Ética (CAAE) n° 87510318.5.0000.5182 e Parecer 2.618.913 (ANEXO I.2).

Capítulo 4

Avaliação Experimental

Neste capítulo, os resultados obtidos são apresentados na busca da base de características fisiológicas, faciais e da atenção visual dos participantes, contemplando-se a análise dos dados obtidos em cada conjunto de ensaios, até aquele que obteve o resultado mais aceitável com os recursos que tinha-se. De posse desses dados, foram também apresentados resultados relacionados à pesquisa de sumarização personalizada de vídeos.

Para a análise dos dados obtidos nos conjuntos de ensaios, foi conduzida uma análise estatística descritiva dos dados sincronizados para cada característica fisiológica do participante em cada vídeo exibido, bem como uma análise do espaço físico onde foi realizada a coleta de dados e dos instrumentos utilizados nesta pesquisa.

4.1 Coleta de Dados

Nesta seção são apresentados os dados de cada conjunto de ensaios. São comentados para cada conjunto de ensaios, os dados e as análises realizadas para avaliação de sua qualidade.

4.1.1 Primeiro Conjunto de Ensaios

Os excertos de vídeos exibidos nestes ensaios foram coletados de vídeos obtidos do YouTube. Os dados foram coletados de modo a respeitar a conveniência dos participantes. Para tanto, foram agendados dias e horários de preferência de cada um deles, bem como organizado o ensaio em seis blocos (um para cada vídeo). Desta forma, caso fosse necessária, a participação poderia ser fracionada e se adequar ao tempo disponível de cada um.

Iniciaram-se os ensaios com treze participantes. Destes, nove completaram os ensaios, ou seja, assistiram aos seis vídeos, tiveram suas reações fisiológicas e expressões faciais aferidas e destacaram os quadros que consideraram mais significativos para composição dos sumários. A exibição dos vídeos que compuseram o ensaio seguiu ordens diferentes para cada participante.

No que se refere às reações fisiológicas durante a realização dos ensaios, aferiu-se e analisou-se o pulsação cardíaca (BPM), a condutância da pele (GSR2), a saturação de oxigênio (SaO2) e a temperatura (TEMPtel).

Duas situações se destacaram para a variável pulsação cardíaca (BPM), quais sejam: em quatro dos seis vídeos exibidos, quatro participantes apresentaram em pelo menos 50% dos casos valores inferiores a 60 bpm ou superiores a 100 bpm, enquanto nos outros dois vídeos analisados, percebeu-se que estes valores de batimentos cardíacos foram registrados em cinco dos nove participantes, assim todos estes estão fora da faixa de normalidade, conforme apresentado na Seção 3.3.2.1, página 31.

Atinente à condutância da pele, os valores mais altos para a GSR2 (diferença relacionada a média da condutância da pele) dos participantes foram registrados durante as exibições dos filmes LOR, BMI e CRA, com valores de 92, 82 e 78 respectivamente.

No tocante à variável Saturação de Oxigênio (SaO2), observou-se que, em pelo menos 75% dos casos, os valores registrados apresentaram-se abaixo de 95%, assim conforme observado na Seção 3.3.2.1, página 31, estes estão fora da normalidade.

Concernente à temperatura dos participantes (TEMPtel), verificou-se que as maiores variações aconteceram na faixa de 1.6° à 4.7°C nos vídeos CRA e FNE. Esta variação de temperatura pode revelar emoções sentidas pelos voluntários, especialmente em dois dos filmes exibidos.

Diante do exposto na Seção 3.3.2.1, página 31, cabe ponderar que, tanto foi possível a temperatura verificada nesta pesquisa ter sofrido aumento em função da exposição dos voluntários às emoções, quanto ter havido falha no processo de aquisição dos dados, a exemplo de posicionamentos inadequados do punho pelo participante.

Analisando-se os registros, observou-se a existência de dados omissos, ou seja, dados que, por algum problema, não foram registrados. Com respeito a esta questão, totalizaram 10 registros, oito deles pertencentes a apenas dois participantes, sendo quatro registros para

cada participante.

A análise criteriosa dos dados coletados no 1° Conjunto de Ensaios, realizado com nove participantes, não demonstrou resultados satisfatórios, uma vez que não estavam fornecendo dados completos e corretos, porém nos permitiu fazer algumas reflexões no tocante à configuração dos experimentos, a saber:

- Os sensores não estavam bem ajustados, apresentando mal contato nos conectores do Arduino e comprometendo às aferições, portanto, gerando medidas não fidedignas e questionáveis;
- Como consequência dos problemas com os sensores, fez-se necessária uma nova coleta com o mesmo participante, que assistiu ao mesmo vídeo por mais de uma vez, comprometendo assim a demonstração das emoções;
- Os vídeos escolhidos nesta fase dos ensaios tinham duração longa, de 10 minutos, o que aumentou a possibilidade de erros nas capturas dos dados fisiológicos; e
- O ambiente onde aconteceu a coleta dos dados, por ser menos controlado, favorecia distrações e desconfortos.

A partir destas reflexões, considerou-se pertinente proceder alguns ajustes e realizar nova coleta de dados (2° Conjunto de Ensaios), de forma a aprimorá-las e conseguir melhores resultados.

4.1.2 Segundo Conjunto de Ensaios

Neste conjunto de ensaios, foram exibidos os vídeos obtidos da Internet, apresentados na base de dados TVSum, os quais foram identificados com siglas, conforme o Quadro 3.1. Realizaram-se coletas com 31 participantes que completaram todas as exigências requeridas por esta pesquisa. No que tange às reações fisiológicas e expressões faciais aferidas durante as coletas, manteve-se as mesmas que foram avaliadas no 1° Conjunto de Ensaios.

Para a análise dos dados sincronizados, estes passaram por um processo de triagem, com vistas a escolher aqueles que se apresentassem mais fidedignos e com nível de ruído aceitável. A referida seleção obedeceu às seguintes etapas de filtragem e refinamento dos registros de dados para cada vídeo e participante:

1. Identificação dos quadros em que os dados fisiológicos do participante não foram registrados;

2. Identificação dos quadros no qual ocorreram dados fisiológicos ruidosos (fora do padrão aceitável); e

3. Identificação dos registros em que a soma total de quadros com dados faltantes e dados ruidosos foi superior a 10% do número total de quadros, este registro foi considerado inválido. Caso contrário, os dados ruidosos são tolerados e os dados faltantes são tratados utilizando interpolação linear.

Os registros pertencentes ao mesmo participante que foram selecionados nesta etapa de filtragem e refinamento, foram organizados e verificados quanto à presença dos três registros dos vídeos da mesma categoria. Em caso afirmativo, estes registros foram utilizados para testar o funcionamento do sumarizador proposto nesta tese. Para tal, os referidos dados foram separados para o treinamento (dois vídeos) e para o teste (um vídeo).

No treinamento, foram gerados sumarizadores personalizados para cada categoria de vídeo e participante. Os registros de dois vídeos e os quadros selecionados pelo participante foram utilizados para treinar um classificador SVM. No teste, o sumarizador foi alimentado com os registros deste novo vídeo e retornou sua seleção de quadros, os quais foram comparados aos quadros selecionados deste novo vídeo pelo participante. Os resultados utilizando as métricas CUS_A e CUS_E foram comparados com a seleção aleatória de quadros.

4.1.2.1 Análise dos Dados Sincronizados

Percebeu-se uma melhora significativa na aquisição dos dados da variável pulso (BPM), pois o percentual de valores considerados fora da normalidade, por faixa de participante e vídeo, variou de 6,4 a 25,8%. No que se referem à condutância da pele (GSR2), os valores mais altos foram registrados durante a exibição dos filmes PK1, FM3 e DS1, com as medidas 184, 174 e 172, respectivamente. Em relação à Saturação de Oxigênio (SaO2), semelhantemente ao 1° Conjunto de Ensaios, mais de 75% dos dados foram considerados abaixo da faixa de normalidade. Sobre a temperatura (TEMPtel) as maiores variações ocorreram durante a exibição dos filmes PK3, BK1 e BK2, respectivamente.

Esta fase de realização das coletas contou com um número ampliado de participantes. Não houve repetição na exibição dos vídeos. Utilizou-se uma base de dados que disponibilizou vídeos com menor duração, na faixa de tempo (1 minuto e 38 segundos a 3 minutos e 55 segundos). Realizaram-se aprimoramentos nos circuitos, com melhor fixação dos fios que os compunham, por intermédio de solda e organização do Arduino. Realizaram-se os ensaios em ambiente controlado (temperatura, iluminação, cores e mínimo barulho, de forma a mitigar as distrações).

No Quadro 4.1, é apresentada a distribuição dos participantes pelos respectivos grupos de vídeos que conseguiram passar pelo filtro utilizado para remover os conjuntos de dados ruidosos.

Quadro 4.1: Apresentação por categorias dos participantes do ciclo de experimentação para o Segundo Conjunto de Ensaios

Categorias	**Participante**
BK	P07, P12, P21, P24, P25, P27, P28
DS	P12, P13, P14
FM	P03, P04, P14, P16, P17, P18, P21, P27
PK	P06, P07, P15, P18, P21, P26

Fonte: **Autor**.

4.1.2.2 Treinamento e Teste

Os dados dos participantes e categorias apresentados no Quadro 4.1 foram utilizados para treinar um classificador SVM, o qual, possibilitou a obtenção dos quadros selecionados pelo sumarizador. O treinamento por categoria do vídeo foi realizado para cada participante individualmente, utilizando-se dois vídeos para treinamento e testando-se com o restante dos vídeos. Na Tabela 4.1, são apresentados os valores da métrica utilizada para realização da avaliação dos sumários. Observa-se que o sumarizador obteve valores semelhantes a uma seleção aleatória dos quadros (que é um indicativo de resultado ruim). Esta informação indicou a ocorrência de problemas relacionados à obtenção dos dados e/ou divergência entre aquilo que foi solicitado ao participante e aquilo que o sumarizador conseguiu realizar com os dados do participante.

Mesmo procedendo-se os ajustes anteriormente descritos, ao final deste conjunto de ensaios e análise dos resultados pertinentes a esta etapa, percebeu-se ainda a necessidade de

mais ajustes e refinamentos, em função de novas observações de possíveis falhas na aquisição dos dados, a saber:

- Foi solicitado ao participante que escolhesse quadros que pudessem resumir os vídeos vistos, que podem não estar relacionados àqueles que mais os emocionaram;
- A temperatura não seria uma reação fisiológica significativa relacionada às emoções, visto que sua variação ao longo de 24 horas é mínima e cíclica (CRAVEN; HIRNLE, 2006);
- A base de dados utilizada poderia conter vídeos que ainda não provocavam reações suficientes para avaliar as emoções estudadas;
- GSR: Posicionamento do sensor, que nesta fase da pesquisa foi colocado nas falanges proximais dos dedos indicador e médio; e
- Rastreador ocular: que demonstrou incompatibilidade com o sistema operacional, necessitando de reinstalação na maioria dos ensaios.

Com base no exposto, considerou-se necessário realizar nova coleta de dados (3° Conjunto de Ensaios), com a correção das possíveis falhas descritas anteriormente, de forma a aprimorar a coleta dos dados e conseguir melhores resultados.

4.1.3 Terceiro Conjunto de Ensaios

Os ajustes e refinamentos para esta nova fase de coletas de dados foi composta basicamente por seis aspectos, quais sejam:

1. Proposição de uma nova base de dados, composta por vídeos que evocassem mais emoções, a qual foi construída com o intuito de pertencer a, pelo menos, um dos gêneros cinematográficos: comédia, horror/thriller, tragédia/drama e neutro;
2. Ajustes para a coleta dos dados:
   (a) Isolamento dos sensores e melhor fixação dos circuitos;
   (b) Remoção do sensor de temperatura (Conforme observado no esquema apresentado no Apêndice C.1);

   (c) Remoção de três resistores do sensor - oxímetro de pulso (MAX30100 - RCWL0530), conforme recomendação do Blog disponível na internet [1];

3. Posicionamento adequado do sensor de condutância da pele GSR, restrito às falanges distais prioritariamente ou falanges mediais do dedo indicador e médio;

4. Reinstalação do sistema operacional (WINDOWS 10);

5. Ajustes necessários nos programas de captura dos dados (Apêndice C.2).

6. A configuração do ensaio sofreu as seguintes alterações:

   (a) Exibição de vídeo relaxante anterior à apresentação do clipe ao participante;

   (b) A quantidade de quadros a serem escolhidos em cada clipe passou a ser variável com número mínimo e máximo de quadros, uma vez que nos conjuntos de ensaios anteriores estas escolhas eram fixas;

   (c) Caso necessário, apresentação de uma sinopse do que aconteceu no vídeo original até o início do clipe exibido para fornecer uma contextualização ao participante.

Neste conjunto de ensaios, os participantes assistiram vídeos de quatro gêneros diferentes obtidos da plataforma YouTube, os quais estavam identificados apenas com as siglas, conforme apresentado no Quadro 3.1, da página 38. As coletas foram realizadas com 33 participantes que cumpriram todos os requisitos para esta pesquisa. A quantidade de participantes foi superior a maioria dos estudos revisados desta área de pesquisa, conforme pode ser observado no Quadro 2.2, da página 21.

Analisando-se as variáveis saturação de oxigênio e pulso, as quais apresentaram percentuais destes valores fora daqueles considerados normais, por participante e vídeo, de 76% e 25,8% respectivamente, no 2° Conjunto de Ensaios, estas evoluíram para a nulidade de ruídos no 3° Conjunto. Desta forma, foi possível adotar os dados destas reações fisiológicas como confiáveis e, portanto, viáveis para os propósitos desta pesquisa. A base de dados obtida, os quadros-chave de todos os vídeos e os *scripts* da avaliação experimental estão disponíveis em repositório do google drive[2].

[1] https://www.teachmemicro.com/max30100-arduino-heart-rate-sensor/
[2] https://drive.google.com/drive/folders/1f95kS-Z4HVpNq6FEwkY27zUm7eWU97wb

Tabela 4.1: Comparação entre o método proposto e o método aleatório na modalidade de encontrar os quadros selecionados por cada participante

Categoria	Participante	Sumarizador*		Sel. Aleatória*	
		CUS_A	CUS_E	CUS_A	CUS_E
BK	P07	0,107	0,893	**0,151**	**0,849**
BK	P12	**0,107**	**0,893**	0,103	0,897
BK	P21	0,085	0,915	**0,151**	**0,849**
BK	P24	**0,229**	**0,771**	0,136	0,864
BK	P25	0,129	0,871	**0,157**	**0,843**
BK	P27	**0,180**	**0,820**	0,139	0,861
BK	P28	**0,250**	**0,750**	0,152	0,848
DS	P12	**0,127**	**0,873**	0,112	0,888
DS	P13	**0,192**	**0,808**	0,160	0,840
DS	P14	**0,157**	**0,843**	0,131	0,869
FM	P03	**0,155**	**0,845**	0,137	0,863
FM	P04	**0,148**	**0,852**	0,145	0,855
FM	P14	**0,186**	**0,814**	0,148	0,852
FM	P16	0,109	0,891	**0,154**	**0,846**
FM	P17	0,128	0,872	**0,144**	**0,856**
FM	P18	0,174	0,826	**0,152**	**0,848**
FM	P21	0,079	0,921	**0,121**	**0,879**
FM	P27	0,072	0,928	**0,138**	**0,862**
PK	P06	0,073	0,927	**0,168**	**0,832**
PK	P07	**0,169**	**0,831**	0,147	0,853
PK	P15	0,086	0,914	**0,169**	**0,831**
PK	P18	0,120	0,880	**0,133**	**0,867**
PK	P21	**0,164**	**0,836**	0,143	0,857
PK	P26	0,128	0,872	**0,160**	**0,840**
Média		**0,140**	**0,860**	**0,144**	**0,856**

* Itens em negrito correspondem aos melhores valores para aquela categoria e participante.
Fonte: **Autor**.

4.2 Investigação da Sumarização Automática de Vídeos usando Características Fisiológicas, Faciais e da Atenção Visual dos Telespectadores

Neste estudo, foi realizada uma análise de algumas estratégias de sumarização de vídeos utilizando-se máquinas de aprendizagem. Para tanto, obedeceu-se à seguinte questão norteadora: **É possível obter sumarizadores automáticos de vídeos a partir de características fisiológicas, faciais e da atenção visual monitoradas dos telespectadores?** A fim de abordar a questão norteadora, foi necessário verificar a eficiência de cada modelo obtido a partir

dos subconjuntos de características comparando a uma seleção de referência.

Abordagens de verificação que medem a magnitude na qual o sumário se sobrepõe a uma seleção de referência é uma abordagem reconhecida e usada em vários estudos no campo da sumarização de vídeos (MONEY; AGIUS, 2010; PENG et al., 2011; OTANI et al., 2019). Otani et al. (2019) argumentaram experimentalmente que o método aleatório experimental produz sumários quase idênticos àqueles apresentados no estado da arte e, até mesmo, àqueles produzidos por anotadores humanos.

Diante destas exposições, formulou-se a seguinte questão secundária: *Quão melhores são os resultados dos sumarizadores automáticos de vídeo obtidos quando comparados aos resultados de sumarizadores aleatórios?*

Para esta investigação, o vetor de características normalizadas ($\check{\Phi}$) (apresentado na Seção 3.6.2, página 50) de cada participante foi reduzido removendo-se as informações referentes aos vídeos classificados no gênero neutro de cada participante e a característica rótulo emocional. Deste novo vetor, foram gerados 20 conjuntos para treinamento e teste. Foi adotada uma proporção usual de 60% e 40% para o treinamento e teste, respectivamente. Todos os conjuntos de treinamentos foram usados em experimentos combinando conjuntos de seis máquinas de aprendizagem com três estratégias de reamostragem. Duas variações de treinamento foram investigadas:

- Utilizando-se todas as características obtidas do participante (TODAS); e
- Utilizando-se apenas as características fisiológicas e expressões faciais (SF).

As máquinas de aprendizagem foram classificadas como regressão logística (LogitBoost), regressão linear (pls e simpls), *naive_bayes*, árvores de decisão (*Random Forest*) e *Support Vector Machines (SVM)*. As seguintes estratégias de reamostragem foram usadas: *Under*, *SMOTE* e *ROSE*.

O método utilizado para a avaliação de cada sumário foi novamente o CUS_A (ver definição na Seção 2.2, página 22). Verificando-se a eficiência dos modelos aprendidos, os comparou com o acaso (Seleção Aleatória). Os quadros auto-reportados de cada participante foram usados como *ground truth* para os sumários. Consequentemente, a seguinte hipótese foi formulada: *CUS_A da seleção aleatória é maior ou igual a CUS_A da seleção da máquina de aprendizagem (com a estratégia de reamostragem apenas no conjunto de treinamento).*

Para testar a hipótese, foi administrado o teste t de Student quando ambas as seleções seguiram a distribuição normal e, em caso contrário, o teste das medianas de Wilcoxon foi administrado. Em qualquer um destes casos, os testes foram executados para amostras pareadas. Quando o p-valor fosse menor que o nível de significância 0,01, a hipótese nula seria rejeitada e a hipótese alternativa seria aceita (*CUS_A da seleção aleatória é menor que CUS_A da seleção da máquina*).

As opções de treinamento SF e TODAS, quando escolhida uma combinação global da máquina de aprendizagem e estratégia de reamostragem (ou seja, a mesma máquina de aprendizagem e estratégia de reamostragem foi adotada para todos os participantes), obtiveram ambas um total de 20 participantes. As combinações que apresentaram o melhor resultado quando comparadas com a seleção aleatória obtiveram 22 e 23 participantes em que esta combinação obteve resultados superiores à seleção aleatória, respectivamente. Nas Tabelas 4.2, 4.3, 4.4 e 4.5, são apresentados os resultados da melhor opção de treinamento (MOT) por participante dos experimentos e a melhor combinação global (MCG), quando existiu para o participante para as opções de treinamento SF e TODAS.

Tabela 4.2: Descrição dos melhores conjuntos de máquinas de aprendizagem e estratégia de reamostragem por participante cujos dados não seguiram uma distribuição normal na opção de treino SF

					Seleção Aleatória		Máquina de Aprendizagem	
ID	Tipo*	Máquina	Reamost.	p.value	CUS_A	$[1^oq; 3^oq.]$	CUS_A	$[1^oq; 3^oq.]$
10	MOT	svmRadial	SMOTE	5.43e-03	0,24	[0,24;0,29]	0,32	[0,29;0,35]
24	MCG	LogitBoost	Under	9.49e-04	0,19	[0,15;0,20]	0,27	[0,22;0,38]
	MOT	svmRadial	SMOTE	3.80e-04			0,31	[0,26;0,31]

Fonte: **Autor**.
* Esta coluna indica a Melhor Opção de Treinamento encontrada (MOT) e o resultado encontrado para a Melhor Combinação Global (MCG) que foi superior ao da seleção aleatória.
Nota: Colunas indicam ID do participante, máquina de aprendizagem, estratégia de reamostragem e medianas correspondentes, 1^o quartil e 3^o quartil para valores CUS_A das máquinas de aprendizagem e a seleção aleatória.

Conforme pode ser visto na coluna "p.value", realçada, das Tabelas 4.2, 4.3, 4.4 e 4.5, essas máquinas obtiveram desempenho significativamente melhor que a seleção aleatória com nível de significância de 0,01 para estes participantes. Assim, pode-se dizer que *"para a maioria dos casos foi possível obter sumários automáticos de vídeos a partir das características fisiológicas, faciais e da atenção visual monitoradas dos telespectadores"*.

Tabela 4.3: Distribuição dos melhores conjuntos de máquinas de aprendizagem e estratégia de reamostragem por participante cujos dados seguiram uma distribuição normal na opção de treino SF

					Seleção Aleatória		Máquina de Aprendizagem	
ID	Tipo*	Máquina	Reamost.	p.value	$\overline{CUS_A}$	Int. Conf.	$\overline{CUS_A}$	Int. Conf.
01	MCG	LogitBoost	Under	4,91e-03	0,30	[0,25;0,34]	0,39	[0,31;0,47]
	MOT	svmRadial	SMOTE	8,85e-04			0,39	[0,34;0,43]
03	MCG	LogitBoost	Under	1,21e-05	0,26	[0,20;0,32]	0,40	[0,33;0,47]
	MOT	rf	SMOTE	1,64e-07			0,42	[0,37;0,46]
04	MOT	svmRadial	SMOTE	9,02e-03	0,24	[0,19;0,29]	0,31	[0,27;0,35]
05	MCG	LogitBoost	Under	1,78e-05	0,26	[0,21;0,31]	0,40	[0,34;0,47]
	MOT	simpls	Under	1,10e-05			0,37	[0,33;0,41]
06	MCG e MOT	LogitBoost	Under	1,91e-04	0,26	[0,22;0,30]	0,33	[0,30;0,37]
07	MCG e MOT	LogitBoost	Under	2,89e-04	0,22	[0,18;0,27]	0,32	[0,26;0,38]
11	MCG	LogitBoost	Under	7,57e-06	0,15	[0,12;0,18]	0,25	[0,21;0,30]
	MOT	rf	SMOTE	6,45e-09			0,30	[0,26;0,34]
12	MCG	LogitBoost	Under	8,77e-05	0,26	[0,22;0,30]	0,39	[0,33;0,45]
	MOT	rf	SMOTE	3,96e-07			0,40	[0,35;0,45]
13	MCG	LogitBoost	Under	1,72e-04	0,17	[0,13;0,22]	0,28	[0,20;0,35]
	MOT	pls	Under	1,42e-04			0,28	[0,23;0,33]
17	MCG	LogitBoost	Under	3,96e-06	0,28	[0,23;0,32]	0,42	[0,36;0,49]
	MOT	pls	Under	5,85e-07			0,40	[0,37;0,42]
19	MCG e MOT	LogitBoost	Under	6,44e-07	0,28	[0,23;0,34]	0,44	[0,40;0,49]
20	MCG e MOT	LogitBoost	Under	4,73e-03	0,19	[0,14;0,23]	0,26	[0,21;0,32]
21	MCG	LogitBoost	Under	3,94e-07	0,23	[0,19;0,28]	0,38	[0,33;0,42]
	MOT	svmRadial	SMOTE	1,09e-08			0,38	[0,34;0,42]
22	MCG	LogitBoost	Under	1,03e-03	0,33	[0,28;0,38]	0,43	[0,38;0,48]
	MOT	pls	Under	5,50e-06			0,46	[0,41;0,50]
23	MCG e MOT	LogitBoost	Under	4,72e-03	0,31	[0,27;0,35]	0,38	[0,33;0,43]
25	MOT	rf	SMOTE	2,29e-05	0,15	[0,10;0,20]	0,24	[0,20;0,28]
26	MCG	LogitBoost	Under	1,97e-05	0,18	[0,12;0,23]	0,34	[0,27;0,41]
	MOT	svmRadial	SMOTE	5,83e-06			0,32	[0,27;0,38]
27	MCG	LogitBoost	Under	4,15e-04	0,19	[0,14;0,23]	0,30	[0,23;0,37]
	MOT	naive_bayes	SMOTE	2,73e-07			0,32	[0,29;0,35]
28	MCG	LogitBoost	Under	6,35e-03	0,21	[0,15;0,27]	0,30	[0,23;0,37]
	MOT	svmRadial	SMOTE	9,96e-06			0,38	[0,33;0,42]
29	MCG e MOT	LogitBoost	Under	1,15e-03	0,23	[0,18;0,27]	0,31	[0,26;0,37]
30	MCG	LogitBoost	Under	2,20e-03	0,21	[0,16;0,26]	0,30	[0,24;0,35]
	MOT	svmRadial	SMOTE	1,15e-05			0,33	[0,30;0,36]

Fonte: **Autor**.

Nota: As colunas contêm ID do participante, máquina de aprendizagem, estratégia de reamostragem e médias correspondentes e intervalos de confiança para os valores CUS_A das máquinas de aprendizagem e seleção aleatória.

* Esta coluna indica a Melhor Opção de Treinamento encontrada (MOT) e o resultado encontrado para a Melhor Combinação Global (MCG) que foi superior ao da seleção aleatória.

Tabela 4.4: Descrição dos melhores conjuntos de máquinas de aprendizagem e estratégia de reamostragem por participante cujos dados não seguiram uma distribuição normal na opção de treino TODAS

					Seleção Aleatória		Máquina de Aprendizagem	
ID	Tipo*	Máquina	Reamost.	p.value	CUS_A	$[1^oq; 3^oq.]$	CUS_A	$[1^oq; 3^oq.]$
10	MCG e MOT	LogitBoost	Under	6,18e-03	0,24	[0,24;0,29]	0,32	[0,26;0,42]
24	MCG	LogitBoost	Under	7,83e-03	0,19	[0,15;0,20]	0,23	[0,22;0,27]
	MOT	rf	ROSE	1,42e-04			0,31	[0,27;0,32]
25	MOT	rf	SMOTE	1,53e-04	0,15	[0,10;0,20]	0,25	[0,20;0,25]

Fonte: **Autor**.
Nota: Colunas indicam ID do participante, máquina de aprendizagem, estratégia de reamostragem e medianas correspondentes, 1^o quartil e 3^o quartil para valores CUS_A das máquinas de aprendizagem e a seleção aleatória.
* Esta coluna indica a Melhor Opção de Treinamento encontrada (MOT) e o resultado encontrado para a Melhor Combinação Global (MCG) que foi superior ao da seleção aleatória.

Os resultados obtidos nas Tabelas 4.2, 4.3, 4.4 e 4.5 podem ser sumarizados como segue:

- Para a maioria dos participantes, foi possível treinar uma máquina de aprendizado (com uma estratégia de reamostragem apenas no conjunto de treinamento) que podia prever (com taxas de acerto e erro melhores do que as decisões aleatórias) se um determinado quadro-chave do vídeo (cujos dados do participante nunca foram vistos pela máquina de aprendizagem) devia ou não fazer parte de um sumário com dimensões pré-determinadas;

- Os resultados referentes à combinação global *versus* específicas (dependentes do participante) de máquina de aprendizagem e estratégia de reamostragem, leva a acreditar que a escolha do par "máquina de aprendizagem" e "estratégia de reamostragem" também deve ser individualizada por participante; e

- A diferença entre a opção de treino TODAS pela a opção SF são os participantes 18 e 32 e a diferença entre a opção de treino SF pela a opção TODAS são os participantes 1, 6 e 20, conforme pode ser observado na Tabela 4.6. Isto indicou que a informação decorrente da fixação ocular para dois participantes foi relevante. No entanto, para a maioria dos participantes, esta foi considerada irrelevante ou, para o caso específico dos três participantes apresentados anteriormente, a fixação ocular piorou os resultados.

Tabela 4.5: Distribuição dos melhores conjuntos de máquinas de aprendizagem e estratégia de reamostragem por participante cujos dados seguiram uma distribuição normal na opção de treino TODAS

					Seleção Aleatória		Máquina de Aprendizagem	
ID	Tipo*	Máquina	Reamost.	p.value	$\overline{CUS_A}$	Int. Conf.	$\overline{CUS_A}$	Int. Conf.
03	MCG	LogitBoost	Under	4.35e-05	0,26	[0,20;0,32]	0,42	[0,34;0,50]
	MOT	rf	Under	1,34e-06			0,42	[0,38;0,47]
04	MCG e MOT	LogitBoost	Under	4,49e-03	0,24	[0,19;0,29]	0,30	[0,26;0,34]
05	MCG e MOT	LogitBoost	Under	3,97e-05	0,26	[0,21;0,31]	0,38	[0,31;0,46]
07	MOT	simpls	Under	7,59e-03	0,22	[0,18;0,27]	0,27	[0,23;0,31]
11	MCG	LogitBoost	Under	4,36e-05	0,15	[0,12;0,18]	0,25	[0,20;0,31]
	MOT	rf	SMOTE	7,47e-08			0,28	[0,23;0,33]
12	MCG	LogitBoost	Under	6,56e-04	0,26	[0,22;0,30]	0,36	[0,31;0,42]
	MOT	rf	Under	3,34e-04			0,36	[0,31;0,41]
13	MCG e MOT	LogitBoost	Under	4,02e-04	0,17	[0,13;0,22]	0,27	[0,21;0,32]
17	MCG e MOT	LogitBoost	Under	1,31e-07	0,28	[0,23;0,32]	0,42	[0,38;0,47]
18	MCG e MOT	LogitBoost	Under	7,62e-03	0,20	[0,15;0,24]	0,28	[0,20;0,35]
19	MCG	LogitBoost	Under	3,89e-06	0,28	[0,23;0,34]	0,42	[0,37;0,47]
	MOT	rf	SMOTE	1,50e-07			0,44	[0,41;0,47]
21	MCG	LogitBoost	Under	2,20e-06	0,23	[0,19;0,28]	0,38	[0,34;0,41]
	MOT	naive_bayes	SMOTE	5,65e-08			0,36	[0,32;0,40]
22	MCG	LogitBoost	Under	4,46e-03	0,33	[0,28;0,38]	0,42	[0,36;0,48]
	MOT	pls	SMOTE	7,85e-05			0,44	[0,40;0,49]
23	MCG e MOT	LogitBoost	Under	5,38e-05	0,31	[0,27;0,35]	0,42	[0,38;0,47]
26	MCG e MOT	LogitBoost	Under	2,01e-05	0,18	[0,12;0,23]	0,31	[0,26;0,36]
27	MCG	LogitBoost	Under	4,81e-04	0,19	[0,14;0,23]	0,28	[0,23;0,32]
	MOT	naive_bayes	ROSE	1,07e-05			0,30	[0,25;0,34]
28	MCG	LogitBoost	Under	2,07e-03	0,21	[0,15;0,27]	0,31	[0,26;0,36]
	MOT	svmRadial	SMOTE	7,58e-05			0,34	[0,29;0,40]
29	MCG e MOT	LogitBoost	Under	5,96e-05	0,23	[0,18;0,27]	0,34	[0,29;0,39]
30	MCG	LogitBoost	Under	9,88e-03	0,21	[0,16;0,26]	0,28	[0,22;0,33]
	MOT	rf	SMOTE	9,09e-04			0,30	[0,26;0,34]
32	MCG e MOT	LogitBoost	Under	9,28e-03	0,27	[0,24;0,31]	0,34	[0,28;0,39]

Fonte: **Autor**.

Nota: As colunas contêm ID do participante, máquina de aprendizagem, estratégia de reamostragem e médias correspondentes e intervalos de confiança para os valores $\overline{CUS_A}$ das máquinas de aprendizagem e seleção aleatória.

* Esta coluna indica a Melhor Opção de Treinamento encontrada (MOT) e o resultado encontrado para a Melhor Combinação Global (MCG) que foi superior ao da seleção aleatória.

Nesta seção, foi realizada uma avaliação objetiva, na qual sumários personalizados automáticos foram comparados com sumários auto-reportados dos correspondentes participantes do experimento. Isso difere do que foi encontrado na literatura da área (PENG et al., 2011; MEHMOOD et al., 2015; LI et al., 2010), onde os participantes são solicitados a dar uma pontuação à sumários automáticos/personalizados depois que eles foram criados. No en-

tanto, os resultados apresentados na Tabela 4.3 e 4.5, corroboram com aqueles obtidos por Money e Agius (2010) em seu estudo que utilizou características fisiológicas para identificar os subsegmentos de vídeo mais divertidos (auto-reportados pelos telespectadores).

Tabela 4.6: Participantes que obtiveram CUS_A melhor que a seleção aleatória.

TODAS	MCG[a]	03, 04, 05, 10, 11, 12, 13, 17, 18, 19, 21, 22, 23, 24, 26, 27, 28, 29, 30, 32
	MOT[b]	03, 04, 05, 07, 10, 11, 12, 13, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32
SF	MCG[a]	01, 03, 05, 06, 07, 11, 12, 13, 17, 19, 20, 21, 22, 23, 24, 26, 27, 28, 29, 30
	MOT[b]	01, 03, 04, 05, 06, 07, 10, 11, 12, 13, 17, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30

Fonte: **Autor**.
[a] MCG = Melhor Combinação Global
[b] MOT = Melhor Opção de Treinamento

4.3 Análise da Relevância das Características Fisiológicas, Faciais e da Atenção Visual dos Telespectadores para a Sumarização Automática de Vídeos Digitais

Com relação à base de dados obtida, a presente seção teve a seguinte questão norteadora: **Haveria um subconjunto reduzido de características fisiológicas, faciais e da atenção visual monitoradas dos telespectadores durante a exibição de vídeos digitais que produzissem sumarizadores automáticos de vídeo?**

A partir do vetor de características normalizadas de cada participante (apresentado na subseção 3.6.2, na página 50) e removendo-se as características obtidas dos vídeos neutros, foram gerados 20 conjuntos de treinamento e teste com uma razão de 60% e 40%, respectivamente. Desta forma, percebe-se uma relação complementar entre os conjuntos de treinamento e de teste, os quais foram utilizados para treinar e testar o classificador *Random Forest* com a estratégia de reamostragem Under, usando-se todas as características.

Posteriormente, empregando-se o seletor de características Boruta, foi obtido o escore de relevância média, conforme Equação 3.7 na página 55, para cada característica dos subconjuntos de características M. Também foi obtido o vetor de características mais relevantes relacionado a um determinado limiar $theta$, conforme Equação 3.9 na página 56, para cons-

trução dos subconjuntos de características S.

Analisando-se os escores da relevância média para cada característica por participante verificou-se que a característica com menor relevância teria $\bar{S}^l$ igual a 0,025 (o seletor indicou "Tentativa" em um experimento e nos demais experimentos, o seletor indicou "Rejeitado") e encontraram-se alguns escores que não foram superiores ou iguais a 0,20, assim determinaram-se essas condições para a construção do conjunto. Nesse sentido, o conjunto completo seria M2_5, M05, M7_5, M10, M12_5 e M15, que são os subconjuntos que identificam as características que obtiveram $\bar{S}^l_k$ maior ou igual a 0,025; 0,05; 0,075; 0,10; 0,125 e 0,15, respectivamente.

Com relação à relevância média normalizada, observou-se que, para alguns participantes, uma única característica tinha este escore superior a 0,25. Desta forma, sistematicamente variou-se θ de 0,30 para 0,80 com passo de 0,025. O objetivo desse procedimento era obter um vetor contendo as características mais relevantes, cuja soma fosse menor ou igual a θ. Utilizando a equação 3.9, encontramos as características mais relevantes para cada θ. Definiu-se S30 e S32_5 quando $\theta = 0,30$ e $\theta = 0,325$, respectivamente e assim por diante para os valores restantes de θ.

As variações dos conjuntos de características descritos anteriormente foram utilizadas para treinar um classificador *Random Forest* para sumarização de vídeos. Os conjuntos de testes foram, então, inseridos nos classificadores treinados para produzir sumários de vídeo. Os sumários obtidos para cada participante foram avaliados usando-se o método conhecido como Comparação de Sumários de Usuários (CUS) (ver definição na Seção 2.2, página 22). Sabendo-se que o número de quadros-chave selecionados para o sumário foi determinado *a priori*, as duas métricas se tornam complementares. Então, foi decidido que os resultados deviam ser avaliados usando apenas os valores de CUS_A. Os valores médios de CUS_A para cada subconjunto de características em cada participante foram calculados.

Por questão de melhor organização do texto, na Tabela 4.7, são apresentadas as médias obtidas para CUS_A e o número médio de características utilizadas para treinamentos e testes em cada subconjuntos de características por participante. O Apêndice H contempla as tabelas apresentando os valores médios do CUS_A obtidos para cada participante em cada subconjunto de características.

Tabela 4.7: Comparação da precisão média das estratégias de seleção de características relevantes e a seleção aleatória para todos os participantes

Grupo	SC	CUS_A	SC	CUS_A	SC	CUS_A	SC	CUS_A	SC	CUS_A	SC	CUS_A
M	M2_5	0,272	M05	0,277	M7_5	0,292	M10	0,296	M12_5	0,298	M15	0,298
S	S30	0,311	S32_5	0,308	S35	0,283	S37_5	0,284	S40	0,286	S42_5	0,287
S	S45	0,290	S47_5	0,287	S50	0,287	S52_5	0,292	S55	0,294	S57_5	0,292
S	S60	0,296	S62_5	0,296	S65	0,298	S67_5	0,299	S70	0,298	S72_5	0,295
S	S75	0,294	S77_5	0,293	S80	0,293					TODAS	0,269
-	Aleat	0,186										

Fonte: **Autor**.

Nota: As colunas apresentam os pares do subconjunto de característica (SC) e o respectivo valor de CUS_A, bem como a qual grupo de subconjuntos pertence. O subconjunto de características TODAS foi colocado no grupo S, porque corresponde ao subconjunto de características cujo somatório é igual a 1. E finalmente, foi apresentado o par da seleção aleatória e o respectivo valor de CUS_A, que não pertence a nenhum grupo de subconjuntos.

A fim de verificar a questão norteadora, foi realizada uma primeiro verificação da eficiência de cada modelo obtido a partir dos subconjuntos de características comparando ao acaso (Seleção Aleatória).

Para criar a seleção aleatória de quadros-chave em cada experimento de treinamento e teste, os quadros-chave foram selecionados aleatoriamente em uma quantidade igual àquela selecionada pelo participante para pertencer ao sumário; assim, os demais quadros-chave foram classificados como não pertencentes ao sumário.

Foi formulada a seguinte hipótese: *o CUS_A da seleção aleatória é maior ou igual ao CUS_A do modelo obtido a partir do subconjunto de características*. Para testar a hipótese foram novamente administrados o teste de comparação das médias t-Student e o teste das medianas de Wilcoxon para amostras pareadas. Quando o p-valor fosse menor que o nível de significância de 0,01, a hipótese nula seria rejeitada e a hipótese alternativa seria aceita (*CUS_A da seleção aleatória é menor que CUS_A do modelo obtido a partir do subconjunto de características*).

Aplicando-se os valores do CUS_A em cada subconjunto de características e os valores obtidos pela seleção aleatória para cada participante, foi realizado o teste de normalidade para estes conjuntos de valores, o que indicou que a maioria seguia a normalidade, exceto os conjuntos do S30 e S32_5. Administrando-se, então, o teste de comparação das medianas de Wilcoxon neste último grupo e o teste de comparação das médias t-Student no primeiro para o teste de hipóteses dos subconjuntos apresentados nas Tabelas 4.8 e 4.9.

Tabela 4.8: Comparação entre diferentes estratégias de seleção de características relevantes por participante e seleção aleatória cujos dados seguiram a distribuição normal

		Seleção de Características		Seleção Aleatória	
	p.value	$\overline{Prec.}$	Int. Conf.	$\overline{Prec.}$	Int. Conf.
M2_5[a]	**3.372e-09**	**0.27**	**[0.23;0.31]**	**0.19**	**[0.17;0.20]**
M05	2,052e-09	0,28	[0,24;0,32]	0,19	[0,17;0,20]
M7_5	1,098e-11	0,29	[0,25;0,33]	0,19	[0,17;0,20]
M10[b]	**1,566e-12**	**0,30**	**[0,26;0,33]**	**0,19**	**[0,17;0,20]**
M12_5	1,015e-11	0,30	[0,26;0,34]	0,19	[0,17;0,20]
M15	8,962e-12	0,30	[0,26;0,34]	0,19	[0,17;0,20]
S35	5,377e-11	0,28	[0,25;0,32]	0,19	[0,17;0,20]
S37_5	8,021e-11	0,28	[0,25;0,32]	0,19	[0,17;0,20]
S40	3,814e-11	0,29	[0,25;0,32]	0,19	[0,17;0,20]
S42_5	2,706e-11	0,29	[0,25;0,32]	0,19	[0,17;0,20]
S45	1,652e-11	0,29	[0,25;0,33]	0,19	[0,17;0,20]
S47_5	3,543e-11	0,29	[0,25;0,32]	0,19	[0,17;0,20]
S50	4,109e-11	0,29	[0,25;0,32]	0,19	[0,17;0,20]
S52_5	1,077e-11	0,29	[0,25;0,33]	0,19	[0,17;0,20]
S55	5,927e-12	0,29	[0,26;0,33]	0,19	[0,17;0,20]
S57_5	9,693e-12	0,29	[0,25;0,33]	0,19	[0,17;0,20]
S60	4,871e-12	0,30	[0,26;0,33]	0,19	[0,17;0,20]
S62_5	9,408e-12	0,30	[0,26;0,33]	0,19	[0,17;0,20]
S65[c]	**3,597e-12**	**0,30**	**[0,26;0,34]**	**0,19**	**[0,17;0,20]**
S67_5	2,183e-11	0,30	[0,26;0,34]	0,19	[0,17;0,20]
S70	1,001e-11	0,30	[0,26;0,34]	0,19	[0,17;0,20]
S72_5	1,235e-11	0,30	[0,26;0,33]	0,19	[0,17;0,20]
S75	1,536e-11	0,29	[0,25;0,33]	0,19	[0,17;0,20]
S77_5	1,071e-11	0,29	[0,25;0,33]	0,19	[0,17;0,20]
S80	1,638e-11	0,29	[0,25;0,33]	0,19	[0,17;0,20]
TODAS	3,536e-09	0,27	[0,23;0,31]	0,19	[0,17;0,20]

Fonte: **Autor**.

Nota: A primeira coluna tem a identificação da estratégia, a segunda coluna contém o p-valor obtido pelo teste t-Student. As colunas restantes apresentaram as precisões médias ($\overline{Prec.}$) E os intervalos de confiança (Int. Conf.) para ambas.

[a] O pior resultado em todos os grupos

[b] Sub-conjunto que obteve os melhores resultados no grupo M

[c] Sub-conjunto que obteve os melhores resultados no grupo S

Os valores obtidos para o p-valor do teste t-Student e do teste de Wilcoxon das Tabelas 4.8 e 4.9, permitiu rejeitar a hipótese nula para todos os subconjuntos de características. Portanto, *CUS_A da seleção aleatória é menor que o CUS_A do modelo obtido de todos os subconjuntos de características apresentados nesta pesquisa.*

No tocante às características mais representativas, foram geradas as Tabelas I.1 e I.2, apresentadas no Apêndice I. A Tabela 4.10 é o resumo destas tabelas, em que são exibidos os subconjuntos M2_5, M10 e S65, que correspondem a pior e melhores subconjuntos,

respectivamente.

Tabela 4.9: Comparação entre diferentes estratégias de seleção de características relevantes por participante e seleção aleatória cujos dados não seguiram a distribuição normal

		Seleção de Características		Seleção Aleatória	
	p.value	Mediana	Quartis	Mediana	Quartis
S30	4,171e-07	0,28	[0,24;0,37]	0,19	[0,15;0,22]
S32_5	4,167e-07	0,28	[0,24;0,37]	0,19	[0,15;0,22]

Fonte: **Autor**.
Nota: A primeira coluna tem a identificação da estratégia, a segunda coluna contém o p-valor obtido pelo teste da mediana de Wilcoxon. As colunas restantes apresentaram a mediana das precisões, 1° e 3° quartis para ambos.

Tabela 4.10: Características consideradas relevantes pelo seletor Boruta nos subconjuntos M2_5, M10 e S65, apresentando-se o número de participantes

			AU*																	Fix*						
	GSR	BPM	01	02	04	05	06	07	09	10	12	14	15	17	20	23	25	26	45	0_5s	1s	1_5s	2s	Fix	NFix	Qu.
M2_5	29	29	30	31	29	23	32	28	30	29	30	30	31	31	28	23	32	30	31	20	23	05	02	19	24	23
M10	19	12	16	19	14	09	22	18	18	17	22	19	14	16	15	10	18	15	14	03	05	00	01	08	05	12
S65	11	06	06	11	04	05	14	12	13	13	15	15	10	09	11	05	11	08	08	00	05	00	01	06	04	11

Fonte: **Autor**.
Nota: AU* e Fix* são os prefixos das características, composta pelo prefixo e substituição do asterisco pelos termos dos rótulos das suas colunas. Qu. corresponde a característica Quadrant (rótulo emocional). Colunas realçadas em verde representam as características mais relevantes para os subconjuntos M10 e S65 e colunas realçadas em vermelho representam as características menos relevantes para o subconjunto M2_5.

Observando-se o subconjunto de características M10, apresentado na Tabela 4.10, que corresponde às características que obtiveram uma relevância igual ou maior que 0,10, temos todas as AU, exceto as AU05 e AU23, todas as respostas fisiológicas e o rótulo emocional foram considerados importantes para pelo menos um terço dos participantes no processo de sumarização.

Em contrapartida, observando-se o subconjunto de características S65, mostrado na Tabela 4.10, em que o somatório de todos os escores de relevância média foi menor ou igual a 0,65, foram considerados relevantes as características AU02, AU06, AU07, AU09, AU10, AU12, AU14, AU20 e AU25, a resposta fisiológica GSR e o rótulo emocional para pelo menos um terço dos participantes no processo de sumarização. Como este conjunto de características é um subconjunto daquele anteriormente obtido, este representa as características mais relevantes encontradas para os participantes.No Quadro 4.2, são apresentadas as AU pertencentes a este subconjunto, a descrição dos movimentos faciais que as caracterizam e os músculos faciais envolvidos. Considerando que:

1. Do conjunto de emoções básicas de Ekman, nojo, raiva, medo, tristeza, alegria, surpresa e desprezo, pelo menos uma destas está relacionada a alguma das AU apresentadas no subconjunto de características mais relevantes, conforme Ekman (1992);

2. Para muitas pesquisas, como Cacioppo, Tassinary e Berntson (2007), o GSR é apresentado como um reflexo de reações fisiológicas que geram excitação, porém estudos como Ayata, Yaslan e Kamasak (2017), apresentam o GSR relacionado tanto a excitação como a valência; e

3. O rótulo emocional apresenta uma relação direta com o estado afetivo do telespectador.

Quadro 4.2: Apresentação das AU mais relevantes encontradas para os participantes

AU	**Descrição**	**Músculo(s) Facial(is) Subjacente(s)**
02	Levantador de Sobrancelha Externa (*Outer Brow Raiser, unilateral, right side*)	*Frontalis (pars lateralis)*
06	Levantador de Bochechas (*Cheek Raiser*)	*Orbicularis oculi (pars orbitalis)*
07	Apertador de Pálpebra (*Lid Tightener*)	*Orbicularis oculi (pars palpebralis)*
09	Enrugador de Nariz (*Nose Wrinkler*)	*Levator labii superioris alaquae nasi*
10	Levantador de Lábio Superior (*Upper Lip Raiser*)	*Levator labii superioris*
12	Puxador de Canto do Lábio (*Lip Corner Puller*)	*Zygomaticus major*
14	Fazedor de Covinhas (*Dimpler*)	*Buccinator*
20	Esticador de Lábio (*Lip Stretcher*)	*Risorius* com *platysma*
25	Separador de Lábios (*Lips part*)	*Depressor labii inferioris* ou relaxamento do *Mentalis*, ou *Orbicularis oris*

Fonte: **Autor**.
Nota: A primeira coluna tem a identificação da unidade de ação facial, a segunda coluna apresenta a descrição dos movimentos dos músculos faciais e a coluna restante contém o(s) músculo(s) facial(is) responsável(is) pelo movimento.

Leva-se a crer que estas características se complementam para a construção do estado afetivo vivenciado em cada momento da exibição do vídeo. Contudo, uma análise mais aprofundada destas características para comprovar esta afirmação é deixada como trabalho futuro.

Ainda foi possível inferir que o subconjunto de características S65, embora tenha sido inicialmente descoberto em uma pesquisa de sumarização personalizada, também pode ser utilizado na obtenção de sumários genéricos.

Para verificar essa generalidade, seguiu-se o mesmo procedimento anterior para a obtenção do CUS_A, exceto pelo fato de que neste experimento o subconjunto de características

utilizado permaneceu o mesmo para todos os participantes, seguindo a escolha previamente definida para o subconjunto genérico para M10 e S65, com as identificações M10G e S65G, respectivamente. Foi formulada a mesma hipótese nula: *o CUS_A da seleção aleatória é maior ou igual ao CUS_A do modelo obtido a partir do subconjunto de características*. Realizou-se o teste de normalidade dos valores de CUS_A, o qual indicou que ambos seguiram a normalidade. Então, administrou-se o teste t-Student para testar a hipótese. Os resultados são mostrados na Tabela 4.11.

Tabela 4.11: Comparação entre as características mais frequentes para os subconjuntos M10 e S65 e a seleção aleatória para todos os participantes

		Seleção de Características		Seleção Aleatória	
	p.value	$\overline{Prec.}$	Int. Conf.	$\overline{Prec.}$	Int. Conf.
M10G	7,946e-09	0,27	[0,23;0,31]	0,19	[0,17;0,20]
S65G	9,378e-09	0,27	[0,23;0,31]	0,19	[0,17;0,20]

Fonte: **Autor**.
Nota: A primeira coluna tem a identificação da estratégia, a segunda coluna contém o p-valor obtido pelo teste t-Student. As colunas restantes apresentaram as precisões médias ($\overline{Prec.}$) e os intervalos de confiança (Int. Conf.) para ambas.

A hipótese nula foi rejeitada pela observação do p-valor do teste t-Student da Tabela 4.11. Então, a hipótese alternativa foi aceita para esses subconjuntos de características. Portanto, *CUS_A da seleção aleatória, novamente, é menor que o CUS_A do modelo obtido desses subconjuntos de características*. Porém, estes subconjuntos de características obtiveram p-valor menos significativos do que os subconjuntos de características apresentados anteriormente nas Tabelas 4.8 e 4.9. Isso evidencia a impossibilidade da construção de conjuntos de características genéricas que obtenham resultados superiores ou iguais àqueles obtidos por um subconjunto de características relacionado a um telespectador para sumarização personalizada.

Na direção oposta, buscando-se as características menos relevantes na Tabela 4.10, analisou-se o subconjunto de características M2_5, que corresponde às características que obtiveram uma relevância igual ou superior a 0,025, ou seja, as características que tiveram alguma relevância. Observou-se que as características Fix1_5s e Fix2s foram consideradas relevantes para poucos participantes (5 e 2 participantes, respectivamente), sendo consideradas, portanto, as menos relevantes. O Fix0_5s e o Fix1s obtiveram um resultado muito melhor. No entanto, Fix1s apresentou ter uma pequena vantagem nos resultados para todos

os subconjuntos de características, exceto o S80, quando comparado com o Fix0_5s, que obteve o mesmo resultado. Contudo, uma análise mais aprofundada deve ser considerada para determinar a melhor duração de fixações para sumarização. Nesse sentido, uma pesquisa para este fim deva considerar durações que variam de 0,5 a 1,5 segundos, sendo esta última descartada.

Seguindo-se a essência da questão norteadora, era necessário fazer uma comparação entre os subconjuntos de características e o conjunto completo. Então, foi formulada a seguinte hipótese nula: *o CUS_A da seleção TODAS é maior ou igual ao CUS_A da seleção do subconjunto*. Conforme observado anteriormente, quase todos os valores seguiram a normalidade, apresentando apenas os subconjuntos S30 e S32_5 que não seguiram. Conforme procedimento realizado anteriormente, administrou-se o teste de comparação das medianas de Wilcoxon e o teste de comparação das médias t-Student para o teste de hipóteses nos subconjuntos em relação à seleção TODAS, que são apresentados nas Tabelas 4.12 e 4.13.

Tabela 4.12: Comparação entre as diferentes estratégias de seleção de características relevantes por participante e a seleção contendo todas as características cujos dados não seguiram a distribuição normal

		Seleção de Características		TODAS	
	p.value	Mediana	Quartis	Mediana	Quartis
S30	1,211e-02	0,28	[0,24;0,37]	0,26	[0,22;0,30]
S32_5	1,985e-02	0,28	[0,24;0,37]	0,26	[0,22;0,30]

Fonte: **Autor**.

Quando observamos os valores obtidos para o p-valor do teste de Wilcoxon e do teste t-Student das Tabelas 4.12 e 4.13, foi possível rejeitar a hipótese nula para a maioria dos subconjuntos, exceto para os casos extremos M2_5, M05, S30, S32_5, S35, S37_5 e S40. Os subconjuntos M2_5 e M05 são das características que apresentaram alguma relevância durante os experimentos e os demais subconjuntos, quando as características correspondentes a 30, 32_5, 35, 37_5 e 40% do escore de relevância média normalizado foram utilizadas. Os resultados para os outros subconjuntos de características permitem rejeitar a hipótese nula, *CUS_A da seleção TODAS é menor que o CUS_A do modelo obtido de todos os subconjuntos de características apresentados nesta pesquisa, exceto para os subconjuntos de características M2_5, M05, S30, S32_5, S35, S37_5 e S40*.

Tabela 4.13: Comparação entre as diferentes estratégias de seleção de características relevantes por participante e a seleção contendo todas as características cujos dados seguiram a distribuição normal

		Seleção de Características		TODAS	
	p.value	$\overline{Prec.}$	Int. Conf.	$\overline{Prec.}$	Int. Conf.
S35	9.420e-02	0.28	[0.25;0.32]	0.27	[0.23;0.31]
S37_5	7.177e-02	0.28	[0.25;0.32]	0.27	[0.23;0.31]
S40	1.376e-02	0.29	[0.25;0.32]	0.27	[0.23;0.31]
S42_5	5.578e-03	0.29	[0.25;0.32]	0.27	[0.23;0.31]
S45	3.134e-03	0.29	[0.25;0.33]	0.27	[0.23;0.31]
S47_5	7.179e-03	0.29	[0.25;0.32]	0.27	[0.23;0.31]
S50	5.017e-03	0.29	[0.25;0.32]	0.27	[0.23;0.31]
S52_5	5.000e-04	0.29	[0.25;0.33]	0.27	[0.23;0.31]
S55	1.558e-04	0.29	[0.26;0.33]	0.27	[0.23;0.31]
S57_5	2.333e-04	0.29	[0.25;0.33]	0.27	[0.23;0.31]
S60	2.146e-05	0.30	[0.26;0.33]	0.27	[0.23;0.31]
S62_5	1.811e-05	0.30	[0.26;0.33]	0.27	[0.23;0.31]
S65	2.893e-06	0.30	[0.26;0.34]	0.27	[0.23;0.31]
S67_5	5.122e-06	0.30	[0.26;0.34]	0.27	[0.23;0.31]
S70	9.468e-07	0.30	[0.26;0.34]	0.27	[0.23;0.31]
S72_5	4.782e-06	0.30	[0.26;0.33]	0.27	[0.23;0.31]
S75	1.185e-05	0.29	[0.25;0.33]	0.27	[0.23;0.31]
S77_5	4.920e-06	0.29	[0.25;0.33]	0.27	[0.23;0.31]
S80	3.192e-06	0.29	[0.25;0.33]	0.27	[0.23;0.31]
M2_5	1.061e-01	0.27	[0.23;0.31]	0.27	[0.23;0.31]
M05	1.845e-02	0.28	[0.24;0.32]	0.27	[0.23;0.31]
M7_5	1.665e-06	0.29	[0.25;0.33]	0.27	[0.23;0.31]
M10	1.552e-07	0.30	[0.26;0.33]	0.27	[0.23;0.31]
M12_5	8.571e-06	0.30	[0.26;0.34]	0.27	[0.23;0.31]
M15	9.405e-05	0.30	[0.26;0.34]	0.27	[0.23;0.31]

Fonte: **Autor**.

4.4 Estudo da Relação entre Expressões Faciais e Estados Emocionais Induzidos pela Apresentação de Conteúdos Multimídia

Nesta seção, foi investigada se expressões faciais apresentadas por telespectadores poderiam estar ligadas às emoções vivenciadas durante exibição de vídeos de diferentes gêneros. Mais especificamente, esta etapa da pesquisa fundamentou-se na seguinte questão norteadora: **Os componentes da expressão facial (AU) podem ser usados para estimar categorias emocionais auto-reportadas durante a exposição a estímulos audiovisuais?** Para melhor análise

e compreensão, fez-se um desdobramento desta questão norteadora nas seguintes questões secundárias: *Pode um subconjunto reduzido de AU produzir melhores resultados de classificação do que aqueles produzidos ao usar todas as AU? Os resultados da classificação pioram quando é restringido o treinamento e os dados de teste a diferentes vídeos?*

As análises utilizadas nesta seção foram traçadas a partir da comparação dos resultados das máquinas de aprendizagem com aqueles de uma seleção aleatória dos quadrantes emocionais atribuídos aos quadros auto-reportados por cada um dos participantes. As *Action Units* e os quadrantes auto-reportados dos participantes do vetor de características normalizadas ($\check{\Phi}$) (apresentado na Seção 3.6.2, página 50) foram usados para treinar cinco diferentes máquinas de aprendizagem, *K-Nearest Neighbor, Support Vector Machine, Random Forest, Neural Network e LogitBoost*. As máquinas que apresentaram melhores acurácias foram consideradas para análise.

Para a primeira avaliação experimental, foram selecionados os vídeos que apresentaram as maiores quantidades de tomadas, rotuladas em cada um dos quadrantes da emoção, para serem utilizadas no treinamento do classificador. Desta forma, tinha-se quatro vídeos no conjunto de treinamento. Os vídeos restantes foram agrupados e somente aqueles que continham rótulos de todos os quatro quadrantes foram usados para formar o conjunto de testes. Os dados dos participantes com vídeos associados que não atenderam aos requisitos supracitados não foram utilizados nos experimentos. Neste sentido, os resultados obtidos de cinco participantes (05, 08, 09, 12 e 26) tiveram que ser removidos da avaliação experimental. Os valores para as acurácias médias, obtidos neste experimento, são mostrados na Tabela 4.14. Os valores destacados nesta tabela são aqueles que estão abaixo da acurácia média de uma rotulagem aleatória de quatro classes (menor que 0,25).

Para comparar a máquina de aprendizagem com a seleção aleatória de quadros-chave, foram realizadas trinta seleções de quadros-chave aleatórios para os rótulos de emoção, e os respectivos valores para acurácia de cada seleção foram calculados. O valor médio para cada participante foi, então, calculado. Os valores da acurácia obtidos ao usar todas as características disponíveis foram, na maioria dos casos, valores superiores aos obtidos pelas acurácias médias da seleção aleatória. Com relação aos participantes, a menor acurácia média obtida foi para o número 18, seguida, em ordem crescente, dos valores das acurácias pelos números 06, 03, 30, 16 e 17, este último com uma acurácia média pouco superior à

0,25.

Tabela 4.14: Acurácias para as máquinas de aprendizagem utilizadas e o modelo base de seleção aleatória, avaliado pelo participante

ID	KNN	SVM	RF	NNET	LB	Aleatória.
01	**0,191**	0,565	**0,243**	0,304	**0,198**	0,248
02	0,352	0,418	0,505	0,462	0,512	0,243
03	**0,264**	**0,250**	**0,181**	**0,236**	**0,148**	0,266
04	**0,195**	0,382	0,333	0,293	0,266	0,243
06	**0,123**	**0,216**	**0,136**	**0,123**	**0,245**	0,254
07	0,275	0,391	0,435	0,406	0,345	0,254
10	0,470	0,554	0,506	0,369	0,476	0,259
11	**0,240**	0,357	**0,170**	0,374	**0,222**	0,241
13	0,542	0,493	0,632	0,458	0,534	0,249
14	0,462	**0,246**	0,415	0,438	0,469	0,256
15	**0,236**	0,377	0,377	0,500	0,348	0,256
16	**0,209**	**0,235**	0,252	**0,243**	0,277	0,245
17	**0,248**	0,271	0,256	**0,241**	**0,244**	0,247
18	**0,070**	**0,174**	**0,243**	**0,122**	**0,221**	0,245
19	0,318	0,268	**0,255**	0,268	0,310	0,256
20	**0,235**	0,452	0,287	0,348	0,307	0,253
21	0,388	0,289	0,579	0,289	0,520	0,253
22	0,492	0,417	0,417	0,492	0,316	0,248
23	0,319	0,391	0,428	0,362	0,431	0,234
24	0,496	0,536	0,504	0,472	0,436	0,231
25	0,435	0,435	0,420	0,442	0,387	0,260
27	0,266	0,287	0,385	0,329	0,267	0,242
28	0,353	0,520	0,529	0,422	0,430	0,239
29	0,343	0,259	0,398	0,481	0,456	0,254
30	0,308	**0,243**	**0,225**	**0,219**	**0,178**	0,253
31	0,421	0,352	0,484	0,415	0,496	0,250
32	0,634	0,611	0,573	0,626	0,591	0,250
33	0,344	0,281	0,312	**0,250**	**0,242**	0,254

Fonte: **Autor**.
Nota: Os valores em negrito indicam os casos em que uma máquina de aprendizagem teve desempenho pior do que a seleção aleatória.

Analisando-se a questão "componentes de expressões faciais (AU) podem ser empregados para classificar categorias emocionais auto-reportadas durante a exposição a estímulos audiovisuais", foi formulada a seguinte hipótese nula: *a seleção aleatória apresenta acurácia maior ou igual à acurácia da máquina de aprendizagem*. Para testar essa hipótese, foram administrados o teste de comparação de média t de Student e o teste das medianas de Wilcoxon conforme exposto anteriormente. Caso a hipótese nula seja rejeitada, a hipótese alternativa seria aceita (*a acurácia da seleção aleatória é menor que a acurácia da seleção*

da máquina).

Utilizando-se os valores originais da acurácia (sem arredondamento) em cada máquina e o valor médio da seleção aleatória para cada participante apresentado na Tabela 4.14, foi realizado o teste de normalidade Anderson-Darling para o nível de significância de 0,01, que indicou que os valores seguem a distribuição normal. Como resultado, foi administrado o teste de comparação das médias t de Student para o teste de hipótese da máquina com o nível de significância de 0,01, conforme apresentado na Tabela 4.15.

Tabela 4.15: Comparação entre as acurácias da seleção aleatória e a máquina de aprendizagem

		Seleção Aleatória		Máquina de Aprendizagem	
	p.value	$\overline{Acuracia}$	Int. Conf.	$\overline{Acuracia}$	Int. Conf.
KNN	1.52e-03	0.25	[0.25;0.25]	0.33	[0.26;0.40]
SVM	1.20e-05	0.25	[0.25;0.25]	0.37	[0.30;0.43]
RF	2.47e-05	0.25	[0.25;0.25]	0.37	[0.30;0.45]
NNET	3.45e-05	0.25	[0.25;0.25]	0.36	[0.29;0.42]
LB	8.33e-05	0.25	[0.25;0.25]	0.35	[0.29;0.42]

Fonte: **Autor**.
Nota: A primeira coluna tem a identificação da máquina de aprendizagem, a segunda coluna contém o p-valor obtido pelo teste t de Student. A acurácia média ($\overline{Acuracia}$) e os intervalos de confiança (Int. Conf.) com nível de significância de 0,01 para ambas as seleções aleatória e máquina de aprendizagem são fornecidos nas colunas restantes.

Ao se observar os valores obtidos para o p-valor do teste t de Student da Tabela 4.15, a hipótese nula foi rejeitada e a hipótese alternativa foi aceita, *a acurácia da seleção aleatória é menor que a acurácia da seleção via máquina de aprendizagem*. Assim, foi possível afirmar que, na maioria dos casos, os componentes da expressão facial (AU) podem ser empregados para classificar as categorias emocionais auto-reportadas durante a exposição aos conteúdos multimídia.

Em um segundo experimento, todas as tomadas de todos os vídeos assistidos pelo participante foram agrupadas: 60% das tomadas foram selecionadas aleatoriamente para treinamento e o restante (40%) para o teste. As máquinas de aprendizagem empregadas foram aquelas que apresentaram melhores resultados (em termos de Acurácia Média, $\overline{Acuracia}$, e Intervalos de Confiança, *Int. Conf.*) no experimento anterior: SVM e *Random Forest*. Este processo foi repetido 30 vezes para evitar qualquer viés. Os valores mínimo e máximo das acurácias obtidas para cada participante são apresentados na Tabela 4.16. As menores acurácias dos participantes estão destacadas em cada linha.

Tabela 4.16: Acurácia mínima e máxima para as máquinas de aprendizagem *Random Forest* e SVM com treinamento e testes usando 60% e 40% de particionamento do conjunto de dados, respectivamente

	Minimo		Máximo	
ID	Random Forest	SVM	Random Forest	SVM
01	0,750	**0,714**	0,920	0,857
02	0,663	**0,564**	0,832	0,752
03	0,344	**0,302**	0,531	0,510
04	0,719	**0,516**	0,859	0,719
05	0,781	**0,646**	0,896	0,844
06	0,757	**0,701**	0,875	0,840
07	0,602	**0,407**	0,707	0,593
08	0,786	**0,753**	0,916	0,909
09	0,855	**0,735**	0,949	0,872
10	0,722	**0,629**	0,844	0,775
11	0,747	**0,714**	0,844	0,851
12	0,661	**0,612**	0,843	0,826
13	0,746	**0,599**	0,894	0,775
14	0,492	**0,467**	0,639	0,672
15	0,767	**0,664**	0,905	0,845
16	0,346	**0,318**	0,514	0,439
17	0,667	**0,598**	0,780	0,735
18	0,415	**0,341**	0,585	0,528
19	0,762	**0,683**	0,884	0,835
20	0,478	**0,433**	0,627	0,590
21	0,630	**0,563**	0,785	0,726
22	0,636	**0,614**	0,750	0,758
23	0,764	**0,655**	0,885	0,811
24	0,662	**0,590**	0,799	0,719
25	0,647	**0,618**	0,824	0,779
26	0,678	**0,653**	0,810	0,843
27	0,779	**0,566**	0,882	0,750
28	0,659	**0,535**	0,791	0,682
29	0,808	**0,733**	0,933	0,933
30	0,588	**0,510**	0,719	0,699
31	0,611	**0,595**	0,825	0,762
32	0,711	**0,648**	0,852	0,796
33	0,623	**0,585**	0,764	0,792

Fonte: **Autor**.

Pode ser visto na Tabela 4.16 que mesmo as menores precisões ainda são maiores que aquelas mostradas na Tabela 4.14 para o mesmo participante. Nesta segunda avaliação, os dados de diferentes quadros do mesmo vídeo podem fazer parte dos conjuntos de treinamento e teste, o que cria melhor distribuição estatística para a decisão do classificador. Observe-se que os dados de treinamento e teste são distintos, porém fortemente relacionados, uma vez que podem vir da mesma exibição de vídeo. Acredita-se que aumentar o número de vídeos no conjunto de dados pode criar uma amostra mais representativa para treinar os modelos e, assim, melhorar os resultados da Tabela 4.14.

Os resultados apresentados nas Tabelas 4.14 e 4.16 foram obtidos usando-se todas as características AU disponíveis como entrada. Seguindo-se a investigação, utilizou-se o seletor de características Boruta, para reduzir o número de AU como entrada para os modelos de aprendizado, obtendo-se o escore da relevância média para cada característica (Equação 3.7

da Seção 4.2, página 55), a diferença consistiu na obtenção do escore após 30 execuções do Boruta e as características analisadas foram apenas as AU.

Para verificar se a redução do número de características melhoraria os classificadores, selecionaram-se as características mais relevantes por participante. Para ser considerado relevante, definiu-se que o escore médio de relevância da característica deveria ser maior ou igual a 0,75. Na Tabela 4.17, são apresentadas as características mais relevantes por participante. As unidades de ação facial 4, 6, 7, 10, 12, 14, 17, 25 e 26 foram consideradas as características mais relevantes para mais da metade dos participantes pelo algoritmo de Boruta.

Tabela 4.17: Características cujo escore de relevância média foi maior ou igual a 0,75

ID	AU01	AU02	AU04	AU05	AU06	AU07	AU09	AU10	AU12	AU14	AU15	AU17	AU20	AU23	AU25	AU26	AU45
01	X	-	X	-	X	X	-	X	X	X	-	X	-	-	-	-	X
02	-	-	X	-	X	X	-	-	-	X	-	-	-	-	-	-	-
03	-	-	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-
04	X	X	X	X	X	X	X	X	X	X	-	X	X	X	-	X	-
05	X	X	X	-	X	X	X	X	X	X	X	X	X	-	X	X	X
06	-	-	X	-	X	X	-	X	X	X	X	X	-	-	-	X	-
07	-	-	X	-	X	X	-	X	X	X	-	-	-	-	X	-	-
08	X	X	X	-	X	X	X	X	X	X	X	X	X	X	X	X	X
09	X	X	X	X	X	-	X	X	-	X	X	X	-	X	X	X	-
10	-	-	X	X	X	X	X	X	X	X	-	X	-	X	-	X	X
11	X	X	X	-	X	X	-	X	X	-	-	X	-	X	-	X	-
12	-	-	X	X	X	X	X	X	X	X	-	-	X	-	X	X	-
13	X	-	X	X	X	X	-	X	X	X	-	X	X	X	X	X	X
14	X	-	-	-	X	X	-	X	X	X	X	X	X	-	X	-	-
15	X	-	X	-	X	X	-	X	X	X	-	-	-	X	X	X	X
16	-	-	-	-	X	-	-	-	-	-	-	-	-	-	-	-	-
17	-	-	X	X	X	X	-	X	X	X	X	-	-	X	-	X	-
18	-	-	X	-	-	-	-	-	-	X	-	-	-	-	X	-	-
19	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
20	-	-	X	-	X	X	X	X	X	X	-	-	-	X	-	-	-
21	X	X	X	-	X	X	-	X	X	X	X	X	-	-	X	X	-
22	-	-	X	-	X	-	-	X	X	X	X	X	X	-	X	X	-
23	X	X	X	-	X	X	-	X	X	X	X	X	-	X	X	-	X
24	X	-	X	X	X	X	-	-	-	X	-	X	-	-	X	X	X
25	X	X	X	-	X	X	-	X	X	X	-	X	X	-	X	-	X
26	-	-	X	X	X	X	-	X	X	X	-	X	-	-	X	-	-
27	-	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
28	X	X	X	X	X	X	X	-	X	-	X	X	-	X	X	X	X
29	-	-	X	-	X	X	-	X	X	X	-	-	-	-	X	X	-
30	-	-	X	-	X	-	-	X	X	X	-	X	-	-	X	-	-
31	-	-	X	-	X	X	X	X	X	X	-	X	X	X	X	X	-
32	X	X	X	-	X	X	X	X	X	X	X	-	X	X	X	-	X
33	-	X	X	-	X	X	-	X	X	X	X	-	-	-	X	X	X
Total	16	13	31	11	31	27	12	27	27	29	14	21	12	15	23	20	14

Fonte: **Autor**.
Nota: (X) Característica com escore de relevância média maior ou igual a 0,75. (-) Característica com escore de relevância média inferior a 0,75.

Todas as características rotuladas como relevantes para cada participante foram apresentadas em um novo experimento usando-se o modelo *Random Forest*. Na Tabela 4.18, são mostradas as acurácias com e sem seleção de características. Nesta Tabela, também são

apresentadas dentre todas as 17 características disponíveis, quantas foram mantidas para o treinamento após a seleção. Os menores valores das acurácias são apresentados destacados.

Verificando-se a questão "um subconjunto reduzido de AU pode produzir melhores resultados de classificação do que aqueles produzidos quando usando todas as AU?", foi formulada a seguinte hipótese nula: *Empregando-se todas as AU nos experimentos obtêm-se uma acurácia maior ou igual àquela obtida ao empregar o subconjunto reduzido de AU*. Para testar esta hipótese, utilizando os valores originais de acurácia (sem arredondamento) apresentados na Tabela 4.18 para cada participante, foi realizado um teste de normalidade, o qual indicou que estes valores seguem a distribuição normal. Em seguida, foi administrado o teste t de Student, que obteve um p-valor = 0,023, acurácias médias e intervalos de confiança de 0,40 ([0,35; 0,44]) e 0,38 ([0,32; 0,43]), para os dois tipos de experimentos por participante, com um subconjunto reduzido de AU e com todas as AU, respectivamente. A partir destes resultados, podemos rejeitar a hipótese nula. Assim, podemos afirmar que, na maioria dos casos, "é possível encontrar um subconjunto reduzido de AU que produza resultados de classificação que sejam superiores aos produzidos quando se utilizam todas as AU".

Nesta seção, foi analisada a ligação entre mudanças nas faces humanas durante a exibição de vídeos e as emoções sentidas. Destacaram-se como resultados desta seção:

- A maioria dos valores de acurácia estava acima daquelas de uma categorização aleatória;
- Conjuntos de treinamento/teste provenientes de vídeos distintos obtiveram valores inferiores de acurácia do que no caso de usar amostras dos mesmos vídeos, a similaridade entre os quadros é reduzida quando se escolhe quadros de vídeos diferentes, levando a expressões faciais mais distintas. Acredita-se que este problema possa ser minimizado apresentando-se mais vídeos aos participantes e, assim, melhorando-se a caracterização dos padrões de emoções a serem aprendidos;
- Forneceu algumas evidências empíricas de que a redução no número de características pode produzir melhor acurácia para a maioria dos participantes; e
- Os resultados apresentados na Tabela 4.16 corroboram com estudos similares como de Wang e Cheong (2006) que usaram recursos audiovisuais para classificar cenas em 36

filmes de Hollywood em 7 rótulos emocionais, assim como de Brezeale e Cook (2006), que avaliou 81 vídeos em gêneros com *closed captions* e coeficientes da transformação de cosseno discreta.

Tabela 4.18: Comparação da acurácia da *Random Forest* para o treinamento e teste com e sem a seleção de características relevantes (SC) por participante

ID	Características reduzidas	Com SC	Sem SC*
01	9	0,383	**0,243**
02	4	**0,495**	0,505
03	1	0,250	**0,181**
04	14	0,342	**0,333**
06	9	0,160	**0,136**
07	7	0,464	**0,435**
10	12	**0,482**	0,506
11	10	0,269	**0,170**
13	14	**0,563**	0,632
14	10	0,438	**0,415**
15	11	0,500	**0,377**
16	1	**0,243**	0,252
17	10	0,256	0,256
18	3	0,261	**0,243**
20	8	0,339	**0,287**
21	12	**0,570**	0,579
22	10	**0,394**	0,417
23	13	0,428	0,428
24	10	**0,480**	0,504
25	12	0,478	**0,420**
27	16	**0,378**	0,385
28	14	0,549	**0,529**
29	8	0,417	**0,398**
30	7	0,225	0,225
31	12	**0,459**	0,484
32	14	0,580	**0,573**
33	11	0,328	**0,312**

Fonte: **Autor**.
* Acurácias foram obtidas usando todas as AU disponíveis (17).

A investigação de rotulagem emocional apresentada levou em consideração vídeos segmentados em tomadas, o que leva a conclusões e aplicações mais refinadas dos resultados, diferentemente de pesquisas anteriores, em que o vídeo inteiro é considerado para fins de rotulagem (por exemplo, Brezeale e Cook (2006), Fischer, Lienhart e Effelsberg (1995), Rasheed, Sheikh e Shah (2005), Huang, Shih e Hsu (2007), Wang et al. (2010) e Hazer et al. (2015)).

Capítulo 5

Considerações Finais

Nesta pesquisa, intentou-se sistematizar uma abordagem de sumarização de vídeos relacionada com a resposta fisiológica e expressões faciais durante a exibição de vídeos. Para tanto, utilizou-se as reações fisiológicas (pulsação e condutância da pele), rastreador ocular e analisador de expressões faciais. Como o método de sumarização de vídeos, apresentado nesta pesquisa, só pode ser realizado com os dados adquiridos dos participantes durante a exibição do vídeo, conclui-se que a abordagem da sumarização é *offline*.

A proposição de um método foi o foco desta tese e para tal, foi necessária uma abordagem incremental que envolveu a realização de múltiplos experimentos, que seguiram direções promissoras e outras não. As promissoras foram aquelas que compuseram as etapas do método proposto.

Considerou-se pertinente retomar às perguntas norteadoras desta pesquisa e destacar os principais pontos que as responderam. Neste sentido, reapresenta-se o primeiro questionamento que nos direcionou: **É possível conceber uma abordagem de sumarização de vídeos baseada, de forma personalizada, em indicadores de emoções humanas?**

As avaliações experimentais apresentadas neste estudo indicam que sumários personalizados produzidos utilizando-se a resposta fisiológica, de atenção visual e unidades de ação facial do telespectador são estatisticamente superiores à sumarização obtida de forma aleatória e sabendo-se que o método aleatório produz sumários semelhantes àqueles apresentados no estado da arte ou àqueles produzidos por anotadores humanos, segundo Otani et al. (2019), consideraram-se viáveis os sumários propostos nesta tese.

Passou-se, então, às ponderações sobre os objetivos específicos que se encarregaram de

responder a este questionamento. Com relação a: *investigar a viabilidade da sumarização automática de vídeos utilizando características fisiológicas, faciais e de atenção visual dos telespectadores*, pode-se inferir que o tipo de sumário proposto nesta pesquisa apresentou-se viável e recomendável, contribuindo inclusive para atender à demanda e exigências da sociedade moderna, que diariamente é "bombardeada" por um expressivo quantitativo de conteúdos digitais, sendo assim imperativa uma ferramenta que auxilie a população a gerenciar esta demanda, uma vez que esta se soma à demanda laboral e às responsabilidades familiares, dentre outras.

Dando prosseguimento às ponderações sobre os objetivos específicos que respondem à esta pergunta norteadora, no que se refere ao objetivo específico: *analisar a relevância das características fisiológicas, faciais e de atenção visual dos telespectadores para o sumário automático de vídeos digitais*. Os resultados obtidos revelaram que os dados fisiológicos e unidades de ação facial dos telespectadores coletados durante a exibição de vídeos são informações valiosas, quando determinada a priori o percentual de sumarização requerido.

Ao analisar as características obtidas dos telespectadores para a construção de sumários automáticos de vídeos digitais, encontraram-se evidências de que as características mais relevantes utilizadas para este fim variam de telespectador para telespectador e ainda que seja possível construir um conjunto reduzido de características personalizado para cada telespectador estatisticamente superior tanto para uma seleção aleatória de quadros, quanto para sumários automáticos construídos utilizando-se todas as características analisadas nesta pesquisa.

Investigando-se a utilização de características fisiológicas, faciais e de atenção visual dos telespectadores para a sumarização automática de vídeos, observa-se que, para a maioria dos participantes, foi possível encontrar pelo menos um par composto de máquina de aprendizagem e estratégia de reamostragem, cujas taxas de acurácia CUS_A obtidas fosse estatisticamente superiores às taxas adquiridas por uma seleção aleatória de quadros-chave para os sumários.

No que se refere ao objetivo específico: *Investigar a relação entre expressões faciais e estados emocionais induzidos pela apresentação de conteúdos multimídia*, na rotulagem de tomadas de vídeos utilizando-se apenas as expressões faciais representadas pelas AU, ficou evidente a superioridade estatística da acurácia dos modelos obtidos das máquinas de

aprendizagem comparada àquela obtida pela classificação aleatória das tomadas.

Cabe destacar ainda que a utilização das AU na construção de sumários foi algo ainda não realizado nas pesquisas revisadas nesta tese. Tal fato reveste esta pesquisa de tese de originalidade e ineditismo, iniciando uma nova direção para pesquisas posteriores.

5.1 Limitações Encontradas

Considerou-se como fatores limitantes para este estudo a carência de pesquisas abordando a temática, em função de ser uma área bastante recente e por este motivo, desafiadora; a necessidade de construir o banco de vídeos que se adequasse melhor a esta proposta, tendo-se em vista que aqueles disponíveis atualmente não demonstraram evocar as emoções no nível que necessitou-se. Além disso, é de suma relevância destacar a necessidade de construir uma sistemática para realizar a sumarização afetiva de vídeos, posto que se trate de um processo pouco explorado e, consequentemente, carente de registros de "como fazer" e que "caminhos seguir".

Além da complexidade de investigar um tema tão subjetivo que é o comportamento humano, outra dificuldade enfrentada ao se realizar este estudo diz respeito ao recrutamento de um número estatisticamente representativo de voluntários para participar dos ensaios.

A sumarização de vídeos com dados adquiridos dos participantes apresentada nesta tese, realiza um sumário do tipo *Storyboard* posteriormente à exibição do vídeo ao participante, que pode ser considerada uma limitação. Entretanto, os quadros obtidos como sumário podem servir como uma opção personalizada dos rótulos utilização para identificação de vídeos em produtos como Netflix ou Sky Play. A forma de obtenção dos dados pode ser considerada outra possível limitação, pois existem restrições relacionadas aos sensores, conforme foi relatado no capítulo de Revisão, porém com os avanços dos vestuários inteligentes e o uso de dispositivos com múltiplas câmeras, essas limitações provavelmente serão minimizadas.

5.2 Publicações Relacionadas a esta Tese

No transcorrer do desenvolvimento desta pesquisa, foram apresentados e aceitos para publicação os artigos científicos: *Investigation of Automatic Video Summarization using Viewer´s*

Physiological, Facial and Attentional Features e *Link between Facial Expression and Emotional State Induced by Exposure to Multimedia Content*, referentes às seções 4.2 e 4.4 e apresentados no APÊNDICE II

É pertinente mencionar que se encontra em fase de elaboração dois trabalhos, um referente à seção 4.3 e outro referente à base de dados, os quais dão prosseguimento e ampliam esta pesquisa.

5.3 Sugestões de Pesquisas Futuras

Como pesquisas futuras, propõem-se realizar novas investigações e aprimorar a abordagem de sumarização proposta nesta tese, conforme discutido nas próximas subseções.

5.3.1 Investigações Futuras

Com base no que foi desenvolvido nesta tese, permite-se sugerir futuras pesquisas que possam trazer contribuição para esta nova área de conhecimento, a saber:

- Expandir o protocolo experimental realizando novos ensaios, considerando a problemática de usuários com problemas oculares como critério de exclusão do participante, assim como ampliando o número de vídeos e participantes para melhor entendimento da relação entre as características dos telespectadores e o estado afetivo dos vídeos;

- Investigar a seleção de características dividindo os participantes em usuário com e sem problemas oculares, considerando restrição existente do rastreador ocular apresentada na pesquisa de Sarmento, Rangel e Gomes (2016);

- Análise semântica das características mais relevantes que emergiram da Seção 4.3 (Análise da Relevância das Características Fisiológicas, Faciais e da Atenção Visual dos Telespectadores para a Sumarização Automática de Vídeos Digitais);

- No tocante ao problema de sumarização, realizar a investigação dos resultados da inclusão do gênero de vídeo na acurácia obtida pelos modelos e, seguindo em outro direcionamento nesta problemática, realizar análises baseadas na idade e gênero dos telespectadores, na tentativa de testar hipóteses mais específicas; e

- Para o problema de rotulagem das tomadas de vídeos, avaliar a adição dos demais dados fisiológicos nos resultados da acurácia dos modelos aprendidos e estudar a relação entre as expressões faciais e dados fisiológicos, bem como a relação entre AU específicas e as precisões obtidas.

5.3.2 Abordagem Proposta para a Sumarização a Partir de Dados Fisiológicos, de Atenção Visual e Unidades de Ação Facial

Pretende-se expandir a abordagem de sumarização proposta nesta pesquisa da seguinte forma. De posse das características de baixo nível do vídeo e do vetor de características normalizadas do usuário, vislumbra-se o treinamento de um classificador que permitirá decidir se um quadro-chave de um vídeo deve ou não pertencer ao sumário do usuário. Identificadas as causas das alterações observadas no vetor de características do usuário, a máquina de aprendizagem torna capaz de aprender os eventos de alto nível e as características de baixo nível do vídeo que dispararam aquelas alterações e replicar esse padrão de eventos em um vídeo que não foi assistido pelo usuário. Para a aprendizagem de máquina na perspectiva supervisionada, foram consideradas duas fases: a fase de treinamento, em que são apresentados os exemplos ao sistema, para que se modifique gradualmente, ajustando seus parâmetros para se aproximar da saída desejada; e a fase de uso, que consiste na apresentação de novos exemplos não vistos anteriormente, visando-se a generalização do sistema.

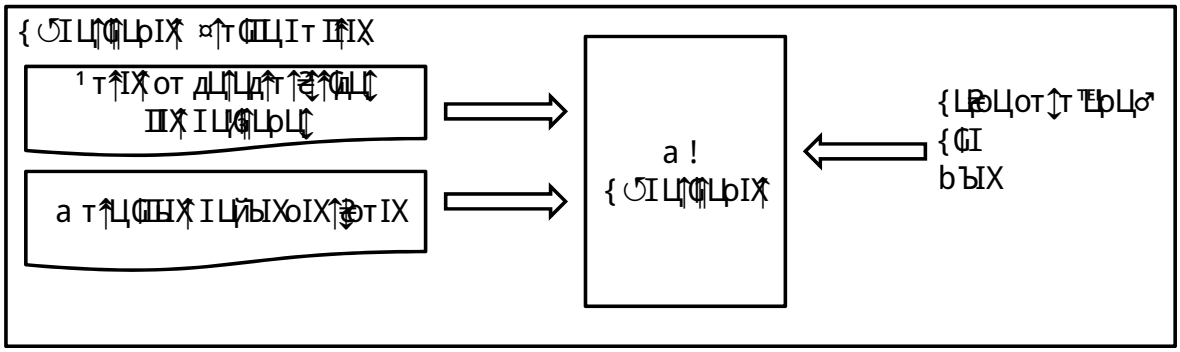
5.3.2.1 Fase de Treinamento

Durante esta fase, acontecerá o treinamento de duas máquinas de aprendizagem conforme apresentado nas Figuras 5.1 e 5.2.

A Figura 5.1 refere-se à máquina de aprendizagem responsável por receber como entrada o vetor de características normalizadas obtidos do telespectador durante a exibição do vídeo referente a cada quadro-chave e as metainformações do vídeo e retornar como saída se o quadro-chave deve pertencer ao sumário ou não.

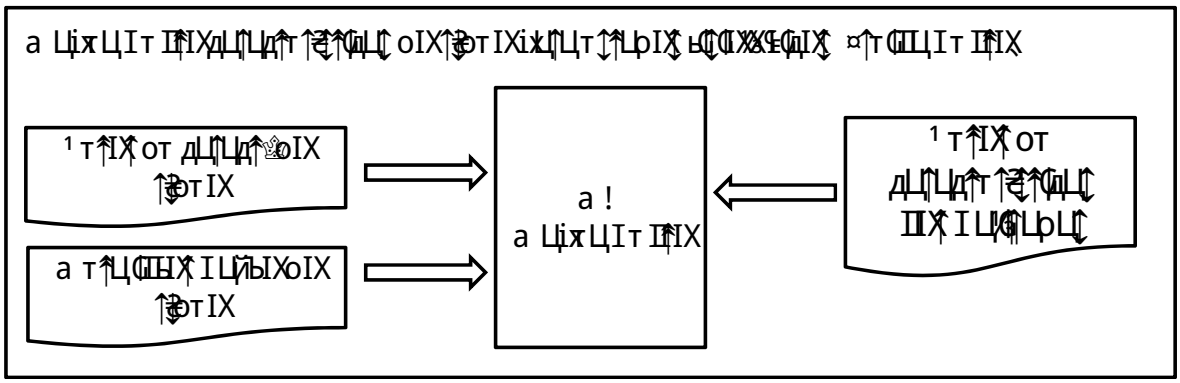
A Figura 5.2 refere-se à máquina de aprendizagem responsável por realizar o mapeamento do vetor de características e as metainformações do vídeo no vetor de características artificiais do telespectador.

Figura 5.1: Fase de treinamento do sumarizador



Fonte: **Autor**.

Figura 5.2: Fase de treinamento para o mapeamento das características do vídeo em características artificiais do telespectador



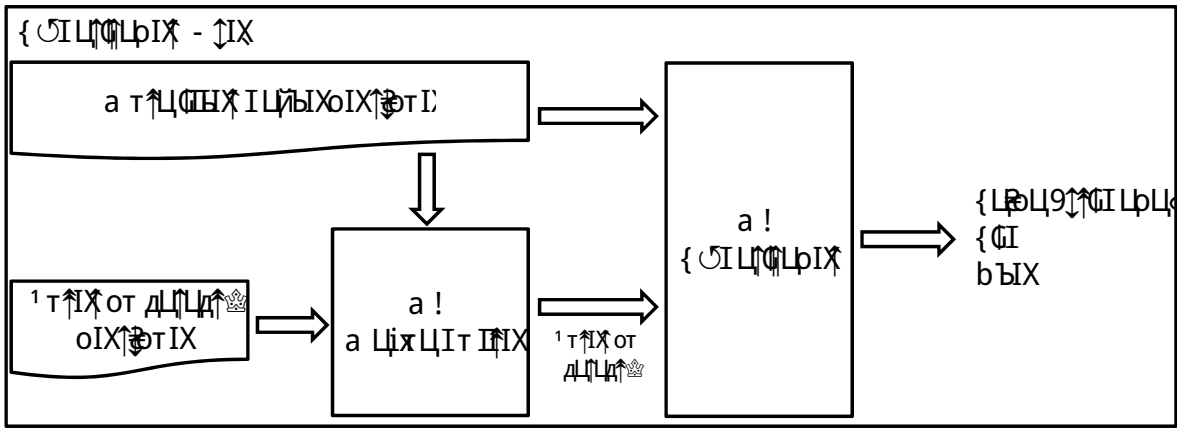
Fonte: **Autor**.

5.3.2.2 Fase de Uso

Na fase de uso, as máquinas de aprendizagem treinadas são utilizadas para executar a sumarização. As máquinas, trabalhando em série, serão capazes de mapear as características de baixo nível de um quadro-chave do vídeo e as metainformações do vídeo e obter como saída se este quadro pertencerá ou não ao sumário. Na Figura 5.3, ilustram-se as etapas do processo de sumarização na utilização.

Neste sentido, destaca-se a relevância deste estudo que no campo acadêmico possibilitará a replicação desta abordagem em outros contextos empíricos, auxiliando outros pesquisadores. Também se ressalta a relevância social ao reduzir, organizar e personalizar o recebimento de informações via vídeos digitais.

Figura 5.3: Fase de uso do sumarizador



Fonte: **Autor**.

Referências Bibliográficas

ALPAYDIN, E. *Introduction to Machine Learning (Adaptive Computation and Machine Learning series)*. Cambridge, MA, USA: The MIT Press, 2009. ISBN 9780262012430.

ARGYLE, M. *Bodily Communication*. [S.l.]: Routledge, 2013.

ATKINSON, L. D.; MURRAY, M. E. *Fundamentos de Enfermagem: Introdução ao Processo de Enfermagem*. Rio de Janeiro, RJ, Brasil: Guanabara Koogan, 2008. ISBN 8527714825.

AURELIO, Y. S. et al. Learning from Imbalanced Data Sets with Weighted Cross-Entropy Function. *Neural Processing Letters*, Springer Science and Business Media LLC, v. 50, n. 2, p. 1937–1949, jan 2019.

AVILA, S. E. F. de et al. VSUMM: A Mechanism Designed to Produce Static Video Summaries and a Novel Evaluation Method. *Pattern Recognition Letters*, Elsevier Science Inc., New York, NY, USA, v. 32, n. 1, p. 56–68, jan 2011. ISSN 0167-8655.

AYATA, D.; YASLAN, Y.; KAMASAK, M. Emotion recognition via galvanic skin response: Comparison of machine learning algorithms and feature extraction methods. *IU-Journal of Electrical & Electronics Engineering*, Istanbul Techinical University, v. 17, p. 3147 – 3156, 2017. ISSN 1303-0914.

BALTRUSAITIS, T.; MAHMOUD, M.; ROBINSON, P. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In: IEEE INTERNATIONAL CONFERENCE AND WORKSHOPS ON AUTOMATIC FACE AND GESTURE RECOGNITION (FG). *Proc.* Ljubljana, Slovenia: Institute of Electrical and Electronics Engineers (IEEE), 2015. v. 11.

BALTRUSAITIS, T.; ROBINSON, P.; MORENCY, L.-P. OpenFace: An open source facial behavior analysis toolkit. In: IEEE WINTER CONFERENCE ON APPLICATIONS OF COMPUTER VISION (WACV). *Proc.* Lake Placid, NY, USA: Institute of Electrical and Electronics Engineers (IEEE), 2016.

BALTRUSAITIS, T. et al. OpenFace 2.0: Facial Behavior Analysis Toolkit. In: IEEE INTERNATIONAL CONFERENCE ON AUTOMATIC FACE AND GESTURE RECOGNITION (FG). *Proc.* Xi'an, China: Institute of Electrical and Electronics Engineers (IEEE), 2018. v. 13.

BASAVARAJAIAH, M.; SHARMA, P. Survey of Compressed Domain Video Summarization Techniques. *ACM Computing Surveys*, Association for Computing Machinery (ACM), v. 52, n. 6, p. 1–29, oct 2019.

BENTO, A. V. Como fazer uma revisão da literatura: Considera[c]ões teóricas e práticas. *Revista JA (Associação Academica da Universidade da Madeira)*, v. 65, p. 42–44, 2012.

BORG, J. *A Arte da Linguagem Corporal - Coleção Vale Mais que Mil Palavras (Em Portuguese do Brasil)*. São Paulo, SP, Brasil: Saraiva, 2012. ISBN 8502138936.

BREIVOLD, H. P.; CRNKOVIC, I.; LARSSON, M. A systematic review of software architecture evolution research. *Information and Software Technology*, Elsevier BV, v. 54, n. 1, p. 16–40, jan 2012.

BREZEALE, D.; COOK, D. J. Using closed captions and visual features to classify movies by genre. In: INTERNATIONAL WORKSHOP ON MULTIMEDIA DATA MINING (MDM). *Proc.* Philadelphia, Pennsylvania, USA, 2006.

CACIOPPO, J. T.; TASSINARY, L. G.; BERNTSON, G. G. (Ed.). *Handbook of Psychophysiology*. Cambridge, United Kingdom: Cambridge University Press, 2007. ISBN 978-0-521-84471-0.

CATAL, C.; DIRI, B. A systematic review of software fault prediction studies. *Expert Systems with Applications*, Elsevier BV, v. 36, n. 4, p. 7346–7354, may 2009.

CELIK, E. A. *Affective analysis of videos: detecting emotional content in real-life scenarios*. Tese (Doctoral Thesis) — University of Berlin, 01 2017. Affective computing; emotional content analysis; violence detection; video content analysis; machine learning; maschinelles Lernen; Videoinhaltsanlayse.

CHAWLA, N. V. et al. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, AI Access Foundation, v. 16, p. 321–357, jun 2002.

CRAVEN, R. F.; HIRNLE, C. J. (Ed.). *Fundamentos de enfermagem: saúde e função humanas*. Rio de Janeiro, RJ, Brasil: Guanabara Koogan, 2006. ISBN 9788527711760.

CRISPIM, A. C. et al. O afeto sob a perspectiva do circumplexo: evidências de validade de construto. *Revista Avaliação Psicológica*, Instituto Brasileiro de Avaliacao Psicologica (IBAP), v. 16, n. 2, p. 145–152, aug 2017.

DAMMAK, M.; WALI, A.; ALIMI, A. M. Video summarization using viewer affective feedback. In: INTERNATIONAL CONFERENCE ON HYBRID INTELLIGENT SYSTEMS, HIS. *Proc.* Gammarth, Tunisia, 2013. p. 279–284.

DAMMAK, M.; WALI, A.; ALIMI, A. M. Viewer's Affective Feedback for Video Summarization. *Journal of Information Processing Systems (JIPS)*, v. 11, n. 1, p. 76–94, 2015.

EJAZ, N. et al. Multi-scale contrast and relative motion-based key frame extraction. *EURASIP Journal on Image and Video Processing*, Springer Science and Business Media LLC, v. 2018, n. 1, jun 2018.

EJAZ, N.; MEHMOOD, I.; BAIK, S. W. Efficient visual attention based framework for extracting key frames from videos. *Signal Processing: Image Communication*, Elsevier Science Inc., New York, NY, USA, v. 28, n. 1, p. 34–44, jan. 2013. ISSN 0923-5965.

EKMAN, P. An argument for basic emotions. *Cognition and Emotion*, Informa UK Limited, v. 6, n. 3-4, p. 169–200, may 1992.

EL-AMIR, S.; EL-FIQI, H. Classification Imbalanced Data Sets: A Survey. *International Journal of Computer Applications*, Foundation of Computer Science, v. 177, n. 23, p. 20–23, dec 2019.

FANG, Y. et al. Visual attention prediction for stereoscopic video by multi-module fully convolutional network. *IEEE Transactions on Image Processing*, Institute of Electrical and Electronics Engineers (IEEE), v. 28, n. 11, p. 5253–5265, nov 2019.

FAROUK, H.; DAHSHAN, K. A. E.; ABOZEID, A. Context-aware joint video summarization and streaming (CVSS) approach. In: IEEE INTERNATIONAL SYMPOSIUM ON MULTIMEDIA (ISM). *Proc.* San Jose, CA, USA: IEEE, 2016.

FELZENSZWALB, P. F. et al. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Institute of Electrical and Electronics Engineers (IEEE), v. 32, n. 9, p. 1627–1645, sep 2010.

FENG, L. et al. Extractive video summarizer with memory augmented neural networks. In: ACM MULTIMEDIA CONFERENCE ON MULTIMEDIA CONFERENCE - MM. *Proc.* Seoul, Republic of Korea: Association for Computing Machinery (ACM), 2018. p. 976–983.

FERREIRA, L.; CRUZ, L. A. da S.; ASSUNÇÃO, P. Towards key-frame extraction methods for 3d video: a review. *EURASIP J. Image and Video Processing*, v. 2016, p. 28, 2016.

FIÃO, G. et al. Automatic generation of sport video highlights based on fan's emotions and content. In: INTERNATIONAL CONFERENCE ON ADVANCES IN COMPUTER ENTERTAINMENT TECHNOLOGY - ACE. *Proc.* Osaka, Japan: ACM Press, 2016.

FISCHER, S.; LIENHART, R.; EFFELSBERG, W. Automatic recognition of film genres. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA (MULTIMEDIA). *Proc.* San Francisco, California, USA: Association for Computing Machinery (ACM), 1995. p. 295–304.

FU, T.-J.; TAI, S.-H.; CHEN, H.-T. Attentive and adversarial learning for video summarization. In: IEEE WINTER CONFERENCE ON APPLICATIONS OF COMPUTER VISION (WACV). *Proc.* Waikoloa Village, HI, USA: Institute of Electrical and Electronics Engineers (IEEE), 2019.

FURINI, M. et al. STIMO: STIll and MOving video storyboard for the web scenario. *Multimedia Tools and Applications*, v. 46, n. 1, p. 47, 2009. ISSN 1573-7721.

GIL, A. C. *Como Elaborar Projetos De Pesquisa (Em Português do Brasil)*. São Paulo, SP, Brasil: Atlas, 2017. ISBN 8597012617.

GUIRONNET, M. et al. Video Summarization Based on Camera Motion and a Subjective Evaluation Method. *EURASIP Journal on Image and Video Processing*, Springer Nature, v. 2007, p. 1–12, 2007.

HALL, T. et al. A systematic literature review on fault prediction performance in software engineering. *IEEE Transactions on Software Engineering*, Institute of Electrical and Electronics Engineers (IEEE), v. 38, n. 6, p. 1276–1304, nov 2012.

HAN, J. et al. Video abstraction based on fMRI-driven visual attention model. *Information Sciences*, v. 281, n. 0, p. 781 – 796, 2014. ISSN 0020-0255. Multimedia Modeling.

HANJALIC, A. *Content-Based Analysis of Digital Video*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2004.

HAZER, D. et al. Emotion Elicitation Using Film Clips: Effect of Age Groups on Movie Choice and Emotion Rating. In: *Communications in Computer and Information Science*. [S.l.]: Springer International Publishing, 2015. p. 110–116.

HE, X. et al. Unsupervised video summarization with attentive conditional generative adversarial networks. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA - MM. *Proc.* Nice, France: ACM Press, 2019.

HILDEBRANDT, A.; OLDERBAK, S.; WILHELM, O. Facial Emotion Expression, Individual Differences in. In: *International Encyclopedia of the Social & Behavioral Sciences*. [S.l.]: Elsevier, 2015. p. 667–675.

HUANG, C.; WANG, H. Novel key-frames selection framework for comprehensive video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, Institute of Electrical and Electronics Engineers (IEEE), p. 1–1, 2019.

HUANG, H.-Y.; SHIH, W.-S.; HSU, W.-H. A Film Classifier Based on Low-level Visual Features. In: IEEE WORKSHOP ON MULTIMEDIA SIGNAL PROCESSING. *Proc.* Crete, Greece: Institute of Electrical and Electronics Engineers (IEEE), 2007. v. 9.

JANG, E.-H. et al. Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. *Journal of Physiological Anthropology*, Springer Nature, v. 34, n. 1, jun 2015.

JI, Z. et al. Video summarization with attention-based encoder-decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, Institute of Electrical and Electronics Engineers (IEEE), p. 1–1, 2019.

JOHO, H. et al. Exploiting facial expressions for affective video summarisation. In: ACM INTERNATIONAL CONFERENCE ON IMAGE AND VIDEO RETRIEVAL, 2009. *Proc.* New York, NY, USA: ACM, 2009. (CIVR '09), p. 31:1–31:8. ISBN 978-1-60558-480-5.

JOHO, H. et al. Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools and Applications*, v. 51, n. 2, p. 505–523, 2011.

K., V. V.; SEN, D.; RAMAN, B. Video skimming: Taxonomy and comprehensive survey. *ACM Computing Surveys*, Association for Computing Machinery (ACM), v. 52, n. 5, p. 1–38, sep 2019.

KATTI, H. et al. Affective video summarization and story board generation using pupillary dilation and eye gaze. In: IEEE INTERNATIONAL SYMPOSIUM ON MULTIMEDIA, 2011. *Proc.* Dana Point, CA, United states, 2011. p. 319 – 326.

KITCHENHAM, B. et al. Systematic literature reviews in software engineering – a systematic literature review. *Information and Software Technology*, Elsevier BV, v. 51, n. 1, p. 7–15, jan 2009.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. [S.l.], 2007.

KOUTRAS, P.; ZLATINSI, A.; MARAGOS, P. Exploring CNN-based architectures for multimodal salient event detection in videos. In: IEEE IMAGE, VIDEO, AND MULTIDIMENSIONAL SIGNAL PROCESSING WORKSHOP (IVMSP). *Proc.* Zagorochoria, Greece: Institute of Electrical and Electronics Engineers (IEEE), 2018.

KUHN, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, Foundation for Open Access Statistic, v. 28, n. 5, 2008.

KURSA, M. B.; JANKOWSKI, A.; RUDNICKI, W. R. Boruta - A System for Feature Selection. *Fundamenta Informaticae*, IOS Press, v. 101, n. 4, p. 271–285, 2010. ISSN 0169-2968.

KURSA, M. B.; RUDNICKI, W. R. Feature Selection with theBorutaPackage. *Journal of Statistical Software*, Foundation for Open Access Statistic, v. 36, n. 11, 2010.

LI, K. et al. Human-centered attention models for video summarization. In: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERFACES AND THE WORKSHOP ON MACHINE LEARNING FOR MULTIMODAL INTERACTION (ICMI-MLMI), 2010. *Proc.* Beijing, China: Association for Computing Machinery (ACM), 2010. p. 27:1–27:8. ISBN 978-1-4503-0414-6.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. *R News*, v. 2, n. 3, p. 18–22, 2002.

LUNARDON, N.; MENARDI, G.; TORELLI, N. ROSE: a Package for Binary Imbalanced Learning. *The R Journal*, The R Foundation, v. 6, n. 1, p. 79, 2014.

MA, Y.-F. et al. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, Institute of Electrical and Electronics Engineers (IEEE), v. 7, n. 5, p. 907–919, Oct 2005. ISSN 1520-9210.

MAAOUI, C.; PRUSKI, A. Emotion recognition through physiological signals for human-machine communication. In: *Cutting Edge Robotics 2010*. [S.l.]: InTech, 2010.

MAVADATI, S. M. et al. DISFA: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, Institute of Electrical and Electronics Engineers (IEEE), v. 4, n. 2, p. 151–160, apr 2013.

MEHMOOD, I. et al. Audio-visual and eeg-based attention modeling for extraction of affective video content. In: INTERNATIONAL CONFERENCE ON PLATFORM TECHNOLOGY AND SERVICE (PLATCON), 2015. *Proc.* Jeju, South Korea: IEEE Computer Society, 2015. p. 17–18.

MEHMOOD, I. et al. Divide-and-conquer based summarization framework for extracting affective video content. *Neurocomputing*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, v. 174, n. A, p. 393–403, 2016. ISSN 0925-2312.

MONEY, A. G.; AGIUS, H. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, v. 19, n. 2, p. 121 – 143, 2008. ISSN 1047-3203.

MONEY, A. G.; AGIUS, H. Analysing user physiological responses for affective video summarisation. *Displays*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, v. 30, n. 2, p. 59 – 70, 2009. ISSN 0141-9382.

MONEY, A. G.; AGIUS, H. ELVIS: Entertainment-led Video Summaries. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, Association for Computing Machinery (ACM), New York, NY, USA, v. 6, n. 3, p. 17:1–17:30, ago. 2010. ISSN 1551-6857.

MONEY, A. G.; AGIUS, H. 'mind the gap': Evaluating user physiological response for multi-genre video summarisation. In: INTERNATIONAL BCS HUMAN COMPUTER INTERACTION CONFERENCE, 27., 2013. *Proc.* Swinton, UK, UK: British Computer Society, 2013. (BCS-HCI '13), p. 37:1–37:6.

MOON, J.-Y. et al. My own clips: Automatic clip generation from watched television programs by using EEG-based user response data. In: IEEE INTERNATIONAL SYMPOSIUM ON CONSUMER ELECTRONICS. *Proc.* Harrisburg, PA, USA: Institute of Electrical and Electronics Engineers (IEEE), 2012.

MUSZYNSKI, M. et al. Aesthetic highlight detection in movies based on synchronization of spectators' reactions. *ACM Transactions on Multimedia Computing, Communications, and Applications*, Association for Computing Machinery (ACM), v. 14, n. 3, p. 1–23, jul 2018.

NGO, C.-W.; MA, Y.-F.; ZHANG, H.-J. Video summarization and scene detection by graph modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, Institute of Electrical and Electronics Engineers (IEEE), v. 15, n. 2, p. 296–305, feb 2005.

NOURBAKHSH, N. et al. Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In: AUSTRALIAN COMPUTER-HUMAN INTERACTION CONFERENCE. *Proc.* Melbourne, Australia: ACM Press, 2012.

OLSON, R. S.; MOORE, J. H. TPOT: A tree-based pipeline optimization tool for automating machine learning. In: *Automated Machine Learning*. [S.l.]: Springer International Publishing, 2019. p. 151–160.

OTANI, M. et al. Rethinking the evaluation of video summaries. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR). *Proc.* Long Beach, CA, United states: Institute of Electrical and Electronics Engineers (IEEE), 2019.

PAIVA, S. C. de; GOMES, H. M. Investigation of automatic video summarization using viewer's physiological, facial and attentional features. In: IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTER COMMUNICATION AND PROCESSING (ICCP). *Proc.* Cluj-Napoca, Romania: IEEE, 2019. No prelo.

PAIVA, S. C. de; GOMES, H. M. Link between facial expressions and emotional states induced by exposure to multimedia content. In: IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTER COMMUNICATION AND PROCESSING (ICCP). *Proc.* Cluj-Napoca, Romania: IEEE, 2019. No prelo.

PAUL, M.; SALEHIN, M. M. Spatial and motion saliency prediction method using eye tracker data for video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, Institute of Electrical and Electronics Engineers (IEEE), v. 29, n. 6, p. 1856–1867, jun 2019.

PENG, W. et al. Editing by Viewing: Automatic Home Video Summarization by Viewing Behavior Analysis. *IEEE Transactions on Multimedia*, v. 13, n. 3, p. 539–550, 2011.

PENG, W.-T. et al. A real-time user interest meter and its applications in home video summarizing. In: IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO. *Proc.* Suntec City, Singapore: IEEE, 2010.

PENG, W.-T. et al. A user experience model for home video summarization. In: *Lecture Notes in Computer Science*. [S.l.]: Springer Berlin Heidelberg, 2009. p. 484–495.

POST, M. J.; PUTTEN, P. van der; RIJN, J. N. van. Does feature selection improve classification? a large scale experiment in OpenML. In: *Lecture Notes in Computer Science*. [S.l.]: Springer International Publishing, 2016. p. 158–170.

PRODANOV, C. C.; FREITAS, E. C. de. *Metodologia do Trabalho Científico: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico - 2ªEdição*. Novo Hamburgo, RS, Brasil: Editora Feevale, 2013.

PRUITT, W. C.; JACOBS, M. Interpreting arterial blood gases: easy as abc. *Nursing*, v. 34, p. 50–53, ago. 2004. ISSN 0360-4039.

QAYYUM, H. et al. Generation of personalized video summaries by detecting viewer's emotion using electroencephalography. *Journal of Visual Communication and Image Representation*, Elsevier BV, v. 65, p. 102672, dec 2019.

RAFF, H.; LEVITZKY, M. G. *Fisiologia Médica: Uma Abordagem Integrada (Lange) (Portuguese Edition)*. [S.l.]: AMGH, 2012. ISBN 9788580551488.

RASHEED, Z.; SHEIKH, Y.; SHAH, M. On the use of computable features for film classification. *IEEE Transactions on Circuits and Systems for Video Technology*, Institute of Electrical and Electronics Engineers (IEEE), v. 15, n. 1, p. 52–64, jan 2005.

ROUT, N.; MISHRA, D.; MALLICK, M. K. Handling Imbalanced Data: A Survey. In: *Advances in Intelligent Systems and Computing*. [S.l.]: Springer Singapore, 2017. p. 431–443.

RUSSELL, J. A. A circumplex model of affect. *Journal of Personality and Social Psychology*, American Psychological Association (APA), v. 39, n. 6, p. 1161–1178, 1980.

RUSSELL, J. A.; BARRETT, L. F. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, American Psychological Association (APA), v. 76, n. 5, p. 805–819, 1999.

SALEHIN, M. M.; PAUL, M. Affective Video Events Summarization Using EMD Decomposed EEG Signals (EDES). In: IEEE CONFERENCE ON DIGITAL IMAGE COMPUTING: TECHNIQUES AND APPLICATIONS (DICTA). *Proc.* Sydney, NSW, Australia: Institute of Electrical and Electronics Engineers (IEEE), 2017.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, IBM, v. 3, n. 3, p. 210–229, jul 1959.

SAQIB, S.; KAZMI, S. Video summarization for sign languages using the median of entropy of mean frames method. *Entropy*, MDPI AG, v. 20, n. 10, p. 748, sep 2018.

SARMENTO, C.; RANGEL, E.; GOMES, H. *Avaliação Comparativa da Usabilidade de Rastreadores Oculares*. [S.l.]: Novas Edições Acadêmicas, 2016. ISBN 3330751150.

SEBASTIAN, T.; J., J. A survey on video summarization techniques. *International Journal of Computer Applications*, Foundation of Computer Science, v. 132, n. 13, p. 30–32, dec 2015.

SINGHAL, A. et al. Summarization of videos by analyzing affective state of the user through crowdsource. *Cognitive Systems Research*, Elsevier BV, v. 52, p. 917–930, dec 2018.

SONG, Y. et al. Tvsum: Summarizing web videos using titles. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR). *Proc.* Boston, MA, USA: Institute of Electrical and Electronics Engineers (IEEE), 2015.

TRUONG, B. T.; VENKATESH, S. Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, Association for Computing Machinery (ACM), v. 3, n. 1, p. 3–es, feb 2007.

WANG, H. L.; CHEONG, L.-F. Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology*, Institute of Electrical and Electronics Engineers (IEEE), v. 16, n. 6, p. 689–704, jun 2006.

WANG, S.; JI, Q. Video Affective Content Analysis: A Survey of State-of-the-Art Methods. *IEEE Transactions on Affective Computing*, Institute of Electrical and Electronics Engineers (IEEE), v. 6, n. 4, p. 410–430, oct 2015.

WANG, W. et al. Revisiting video saliency: A large-scale benchmark and a new model. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.]: IEEE, 2018.

WANG, Z. et al. YouTubeCat: Learning to categorize wild web videos. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. *Proc.* San Francisco, CA, USA: Institute of Electrical and Electronics Engineers (IEEE), 2010.

WAZLAWICK, R. *Metodologia de Pesquisa Para Ciência da Computação*. Rio de Janeiro, RJ, Brasil: Elsevier, 2014. ISBN 853527782X.

WESTERINK, J. H. D. M. et al. Computing emotion awareness through galvanic skin response and facial electromyography. In: *Probing Experience*. [S.l.]: Springer Netherlands, 2008. p. 149–162.

XIANG, X.; KANKANHALLI, M. S. Affect-based Adaptive Presentation of Home Videos. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA (MM), 19. *Proc.* Scottsdale, Arizona, USA: Association for Computing Machinery (ACM), 2011. p. 553–562. ISBN 978-1-4503-0616-4.

XU, M. et al. Hierarchical movie affective content analysis based on arousal and valence features. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA (MM), 16. *Proc.* Vancouver, British Columbia, Canada: Association for Computing Machinery (ACM), 2008. p. 677–680. ISBN 978-1-60558-303-7.

YIN, Y.; THAPLIYA, R.; ZIMMERMANN, R. Encoded semantic tree for automatic user profiling applied to personalized video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, Institute of Electrical and Electronics Engineers (IEEE), v. 28, n. 1, p. 181–192, jan 2018.

YOSHITAKA, A.; SAWADA, K. Personalized video summarization based on behavior of viewer. In: INTERNATIONAL CONFERENCE ON SIGNAL IMAGE TECHNOLOGY AND INTERNET BASED SYSTEMS. *Proc.* Naples, Italy: IEEE, 2012.

YOU, J. et al. A multiple visual models based perceptive analysis framework for multilevel video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, Institute of Electrical and Electronics Engineers (IEEE), v. 17, n. 3, p. 273–285, mar 2007.

ZAWBAA, H. M. et al. Event detection based approach for soccer video summarization using machine learning. *International Journal of Multimedia and Ubiquitous Engineering (IJMUE)*, v. 7, n. 2, p. 63–80, 2012.

ZHANG, L. et al. Guest editors' introduction: Perception, aesthetics, and emotion in multimedia quality modeling. *IEEE MultiMedia*, v. 23, n. 3, p. 20–22, July 2016.

ZHANG, Y.; TAO, R.; WANG, Y. Motion-state-adaptive video summarization via spatiotemporal analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, Institute of Electrical and Electronics Engineers (IEEE), v. 27, n. 6, p. 1340–1352, jun 2017.

ZHANG, Z. et al. Show, tell and summarize: Dense video captioning using visual cue aided sentence summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, Institute of Electrical and Electronics Engineers (IEEE), p. 1–1, 2019.

ZHAO, B.; LI, X.; LU, X. HSA-RNN: Hierarchical structure-adaptive RNN for video summarization. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. *Proc.* Salt Lake City, UT, USA: Institute of Electrical and Electronics Engineers (IEEE), 2018.

Apêndice A

Revisão Sistemática

De acordo com Bento (2012), a pesquisa bibliográfica ou revisão de literatura é uma etapa imprescindível a investigação científica, a qual trará influência em todas as etapas posteriores. Com esta compreensão, foi realizada uma revisão sistemática da literatura que se norteou pelas seguintes questões: **De que formas (categorias de métodos, modalidades percebidas) a sumarização automática de vídeos, com base no monitoramento do usuário (ou da audiência), tem sido realizada? Como foram conduzidas as avaliações experimentais da literatura revisada (categorização, objetivas ou subjetivas)? Quais as bases de dados empregadas?**

A.1 Metodologia Adotada para a Revisão da Literatura

Na revisão sistemática realizada neste estudo, adotou-se a metodologia de pesquisa proposta por Kitchenham e Charters (2007). Por este motivo, adotaram-se as etapas descritas a seguir:

A.1.1 Termos de Busca

A partir da definição das questões de pesquisa, foi possível identificar um conjunto de termos de busca. Buscou-se em seguida, ampliar esse conjunto com sinônimos. A expressão de busca foi criada combinando-se os termos de busca e seus sinônimos. Tal expressão lógica foi previamente agrupada dentro de cada linha de pesquisa segundo seu significado. O operador booleano OR foi utilizado para conectar os sinônimos identificados, enquanto que o operador booleano AND foi empregado para conectar os grupos de termos de busca.

Utilizaram-se os termos de busca em inglês, por ser uma língua de maior abrangência e concentrar a maioria das bases de dados bibliográficas digitais relacionadas à temática. Este agrupamento pode ser visto no Quadro A.1.

Quadro A.1: Termos de busca agrupados segundo o significado semântico ou relacionados ao mesmo domínio

	Termos/Sinônimos
Grupo 1	*affect*; *attention*; *emotion*
Grupo 2	*summarization*
Grupo 3	*video*; *clip*; *movie*

Fonte: **Autor**.

A.1.2 Seleção das Fontes de Pesquisa

A pesquisa foi realizada mediante buscas eletrônicas de estudos científicos, condizentes com o objetivo proposto neste estudo, disponibilizados em conferências, simpósios e revistas publicadas no período de janeiro de 2005 a outubro de 2019.

As bases digitais foram exploradas por intermédio dos seus sítios oficiais, com o auxilio dos respectivos engenhos de busca. Utilizou-se a mesma expressão lógica de busca para todas as bibliotecas digitais acessadas, apenas realizando adaptações para ajustarmos ao padrão adotado pelo engenho de busca específico de cada base dados bibliográfica. A seleção das bases de dados seguiu os seguintes critérios:

- Disponibilidade das publicações nas bases de dados bibliográficas;
- Disponibilidade de busca a partir de termos de buscas e expressões lógicas;
- Relevância da base de dados bibliográfica; e
- Abrangência de veículos de publicação mais importantes e com os maiores impactos na comunidade científica.

As bases de dados bibliográficas digitais encontradas que atenderam aos critérios de seleção anteriormente apresentados foram as seguintes:

- Engineering Village [1];

[1] http://www.engineeringvillage.com/search/quick.url

- IEEE Xplore [2];
- Science Direct [3]; e
- Scopus [4].

Após a identificação das bases de dados bibliográficas digitais, foram selecionadas as publicações a partir das expressões de busca nos respectivos engenhos de busca (Quadro A.2).

Quadro A.2: Base de dados e respectiva expressão de busca

Engineering Village
((((((“*attention*” OR “*affect**” OR “*emotion**”) AND (“*summarization*”) AND (“*video*” OR “*clip*” OR “*movie*”)))) WN ALL)) AND ((2019 OR 2018 OR 2017 OR 2016 OR 2015 OR 2014 OR 2013 OR 2012 OR 2011 OR 2010 OR 2009 OR 2008 OR 2007 OR 2006 OR 2005) WN YR)
IEEE Xplore*
((*attention* OR *affect** OR *emotion**) AND (*summarization*) AND (*video* OR *clip* OR *movie*))
Science Direct *
(“*attention*” OR “*affect*” OR “*emotion*”) AND (“*summarization*”) AND (“*video*” OR “*clip*” OR “*movie*”))
Scopus
TITLE-ABS-KEY ((“*affect**” OR “*attention*” OR “*emotion**”) AND (“*summarization*”) AND (“*video*” OR “*clip*” OR “*movie*”)) AND PUBYEAR > 2004 AND PUBYEAR < 2020

* A filtragem relacionada aos anos foi realizada diretamente no site de busca da base de dados bibliográfica e para estas bases não aparecem na expressão de busca.
Fonte: **Autor**.

A.1.3 Seleção das Publicações

A seleção obedeceu à seguinte sequência:

1. Seleção inicial a partir das expressões de busca;
2. Leitura do título das publicações selecionadas inicialmente;
3. As publicações com títulos sem relação com o objetivo de nosso estudo foram descartados (rejeição na fase de Seleção);

[2] http://ieeexplore.ieee.org
[3] http://www.sciencedirect.com
[4] http://scopus.com

4. As publicações que apresentaram relação clara ou duvidosa com o objetivo do nosso estudo seguiram no processo de avaliação com a leitura dos seus respectivos resumos;

5. As publicações cujos resumos não demonstraram relação com o objetivo de nosso estudo foram descartadas (rejeição na fase de Extração);

6. As publicações cujos resumos apresentaram relação clara ou duvidosa seguiram no processo de avaliação com a leitura da introdução e conclusões; e

7. Após a leitura das introduções e conclusões, as publicações que não demonstraram forte relação com o objetivo de nosso estudo foram descartadas e as demais seguiram para a leitura na íntegra. Os que permaneceram em consonância com o objetivo proposto foram eleitos para participarem da RS e os demais foram descartados.

No que se refere aos critérios de inclusão e exclusão aplicados às publicações contidas nesta RS, foram descritos no Quadro A.3.

Quadro A.3: Critérios de inclusão e exclusão empregados na revisão bibliográfica

Critérios de Inclusão
Artigos que possuem relevância com relação à pergunta de pesquisa. Assim, apenas serão incluídos artigos que descrevam pesquisas relacionadas ao tema sumarização de vídeos a partir de análise afetiva do telespectador.
Artigos que analisem bases de dados relacionadas ao tema de pesquisa.
Artigos que analisem métricas de software (acurácia, performance entre outras).
Artigos que analisem técnicas de aprendizagem de máquina (identificação e classificação) e mecanismos para avaliação de desempenho para tais técnicas de aprendizagem.
Artigos que foram publicados em conferências, simpósios, periódicos e relatórios técnicos.
Artigos publicados no idioma inglês.
Critérios de Exclusão
Artigos que não possuem relevância com relação à pergunta de pesquisa.
Artigos que tratam de sumarização egocentrica e de vigilância
Artigos que não incluem resultados experimentais.
Artigos que possuam conteúdo incompleto.
Artigos que foram publicados em editoriais, prefácios, artigos de resumos, entrevistas, notícias, revisões, cartas, discussões, comentários, tutoriais, *workshops*, painéis e pôsteres.
Artigos não publicados no idioma inglês.
Artigos que não estejam disponíveis online.

Fonte: **Autor**.

As buscas nas bases bibliográficas foram realizadas no dia 21 de outubro de 2019. O propósito das expressões de busca é permitir reprodutibilidade da pesquisa bibliográfica. Não obstante, é possível que os resultados obtidos em uma data futura se alterem devido a

diversos fatores, tais como: alterações nas formas de indexação ou na sintaxe das expressões de busca das bases bibliográficas, bem como a inclusão tardia de artigos em tais bases.

O processo de pesquisa nas bases bibliográficas retornou um total de 629 publicações como resultado. Após a etapa de seleção, verificou-se que 312 eram duplicados (relacionados em mais de um engenho de busca) e 221 foram rejeitados (não se relacionavam ao escopo da pesquisa). Restaram 96 aceitos para a fase de extração. Na fase de extração, 78 publicações foram rejeitadas, restando 18 publicações. Os dados relacionados aos engenhos de busca são resumidos na Tabela A.1.

Tabela A.1: Etapas do processo de revisão e quantidade de artigos selecionados

	Seleção				**Extração**	
Engenho de Busca	Total	Duplicados	Rejeitados	Aceitos	Rejeitados	Aceitos
Engineering Village	194	192	002	000	000	000
IEEE	112	000	076	036	029	007
Science Direct	053	017	034	002	000	002
Scopus	270	103	109	058	049	009
Total	629	312	221	096	078	018

Fonte: **Autor**.

Durante a realização da seleção, os artigos duplicados foram classificados nos engenhos de busca a partir da seguinte ordem de prioridade do maior para o menor: *IEEE*, *Scopus*, *Science Direct*, *Engineering Village*. Por exemplo, supondo-se que um mesmo artigo aparecesse em dois ou mais engenhos, se um desses engenhos fosse o *IEEE*, tal artigo seria então contabilizado o referido engenho de busca. Os valores que aparecem em rejeitados e aceitos na Tabela A.1, são referentes aos totais de aceitos na fase de seleção.

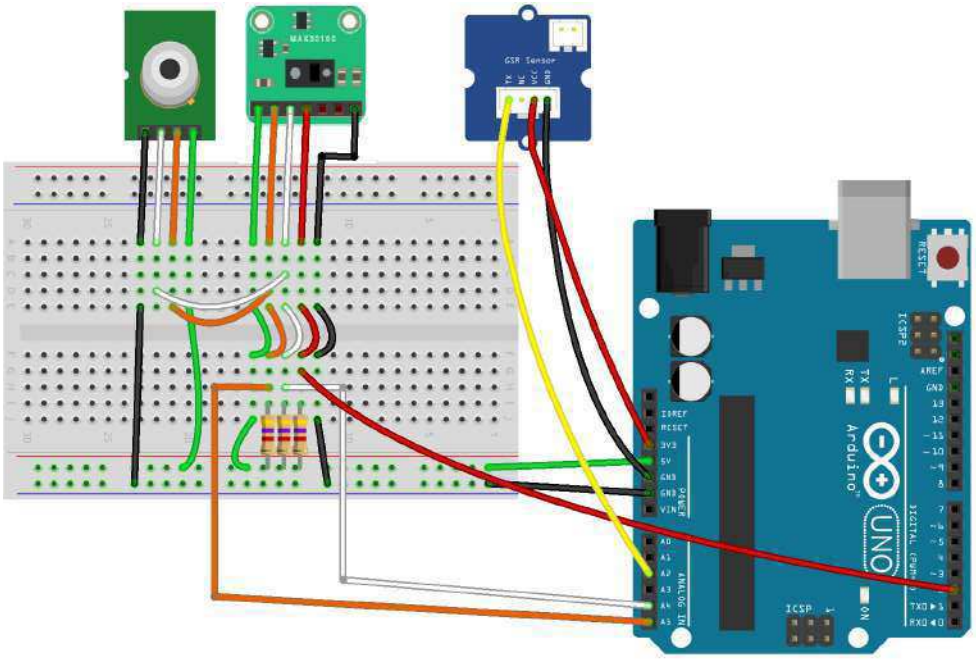
Apêndice B

Arduino - Primeiro e Segundo Conjuntos de Ensaios

Neste apêndice, apresentamos o esquema gráfico da montagem do Arduino e a codificação utilizada no Primeiro e Segundo Conjuntos de Ensaios.

B.1 Montagem do Arduino

Figura B.1: Esquema de montagem do Arduino - primeiro e segundo conjunto de ensaios



Fonte: **Autor**.

B.2 Codificação Utilizada no Arduino

```
#include <Arduino.h>
#include <math.h>
#include <Wire.h>
#include <SparkFunMLX90614.h>
#include "MAX30100.h"
#define PERIOD_OXI 10000 // period in us, periodo = 10^4  x 10^-6 = 10^-2 = 0,01. Frequencia = 1/0,01  = 100Hz
#define PERIOD_GSR_TEMP 5000
unsigned long last_us_OXI = 0L;
unsigned long last_us_GSR_TEMP = 0L;

IRTherm therm;

MAX30100* pulseOxymeter;
const int GSR = A2;
int threshold = 0;
int sensorValueGSR;
float tempAmbiente, tempTelespectador;

void setup(){
  long sum = 0;
  Wire.begin();
  Serial.begin(115200);
  therm.begin();       //Inicializa sensor de temperatura infravermelho
  therm.setUnit(TEMP_C);  //Seleciona temperatura em Celsius
  pulseOxymeter = new MAX30100( DEFAULT_OPERATING_MODE, MAX30100_SAMPLING_RATE_200HZ,  MAX30100_PULSE_WIDTH_400US_ADC_14, DEFAULT_IR_LED_CURRENT, true, true);
  pinMode(2, OUTPUT);
  for (int i = 0; i < 500; i++){
    sensorValueGSR = analogRead(GSR);
    sum += sensorValueGSR;
    delay(5);
  }
  threshold = sum / 500;
}

void loop(){
  int gsr, mudou;
  if (micros() - last_us_OXI > PERIOD_OXI){
    last_us_OXI += PERIOD_OXI;
    if (micros() - last_us_GSR_TEMP > PERIOD_GSR_TEMP){
      last_us_GSR_TEMP += PERIOD_GSR_TEMP;
      if (therm.read()){
        tempAmbiente = therm.ambient();
        tempTelespectador = therm.object();
      }
      sensorValueGSR = analogRead(GSR);
      gsr = threshold - sensorValueGSR;
      if (abs(gsr) > 60){
        sensorValueGSR = analogRead(GSR);
        gsr = threshold - sensorValueGSR;
        if (abs(gsr) > 60) {
          mudou = 1;
        } else {
          mudou = 0;
        }
      } else {
        mudou = 0;
      }
```

```
58      }
59      //You have to call update with frequency at least 37Hz. But the
        closer you call it to 100Hz the better, the filter will work.
60      pulseoxymeter_t result = pulseOxymeter->update();
61      if ( result.pulseDetected == true){
62        if (mudou == 1) {
63          imprimeGSR("{GSR;1;", gsr);
64        } else {
65          imprimeGSR("{GSR;0;", gsr);
66        }
67        Serial.print(";");
68        imprimeTemperatura("TEMPamb;", tempAmbiente , ";TEMPtel;",
        tempTelespectador);
69        Serial.print(";");
70        imprimeOximetro("BPM;", result.heartBPM , ";SaO2;", result.SaO2);
71        Serial.println("}");
72      }
73    }
74  }
75
76  void imprimeGSR(char * leftStr, int MyVar){
77    Serial.print(leftStr);
78    Serial.print(MyVar);
79  }
80
81  void imprimeTemperatura(char *leftStr, float MyVar, char *sepStr, float
        MyVar1){
82    Serial.print(leftStr);
83    Serial.print(MyVar);
84    Serial.print(sepStr);
85    Serial.print(MyVar1);
86  }
87
88  void imprimeOximetro(char *leftStr, float MyVar, char *sepStr, float
        MyVar1){
89    Serial.print(leftStr);
90    Serial.print(MyVar);
91    Serial.print(sepStr);
92    Serial.print(MyVar1);
93  }
94
```

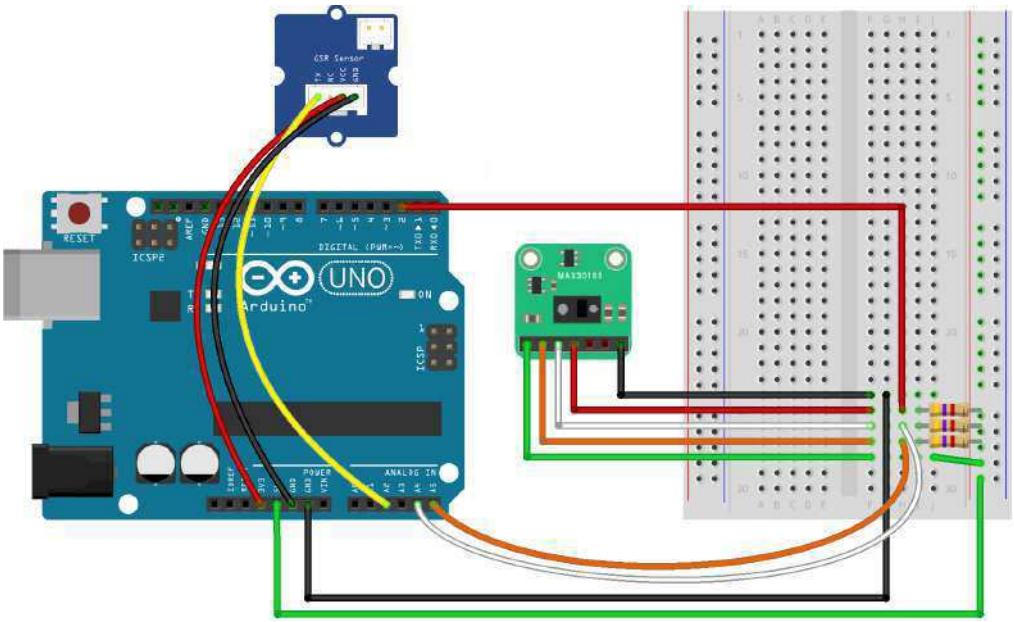
Apêndice C

Arduino - Terceiro Conjunto de Ensaios

Neste apêndice, expôs-se o esquema gráfico da montagem do Arduino e a codificação utilizada no Terceiro Conjunto de Ensaios.

C.1 Montagem do Arduino

Figura C.1: Esquema de montagem do Arduino - terceiro conjunto de ensaios



Fonte: **Autor**.

C.2 Codificação Utilizada no Arduino

```
#include <Arduino.h>
#include <math.h>
#include <Wire.h>
#include "MAX30100_PulseOximeter.h"

#define REPORTING_PERIOD_MS     400
PulseOximeter pox;
#define BUFFER_SIZE 10
int speakerPin = 8;//BUZZER
uint32_t tsLastReport = 0;
unsigned long timefinal[BUFFER_SIZE+1], time2;
double bpm_aux;

void insertshift(unsigned long novot){
  for(int i=0; i<BUFFER_SIZE;i++)
    timefinal[i]=timefinal[i+1];
    timefinal[BUFFER_SIZE] = novot;
}
void calcbpmaux(){
  bpm_aux = BUFFER_SIZE*60000000.0/(timefinal[BUFFER_SIZE]-timefinal[0]);
    //
}

const int GSR = A2;
int threshold = 0;
int sensorValueGSR;

// Callback (registered below) fired when a pulse is detected
void onBeatDetected(){
  time2=micros();
  insertshift(time2);
  calcbpmaux();
}

void setup(){
  long sum = 0;
  Wire.begin();
  Serial.begin(115200);
  pinMode (speakerPin, OUTPUT);//BUZZER
  for (int i = 0; i < 500; i++){
    sensorValueGSR = analogRead(GSR);
    sum += sensorValueGSR;
    delay(5);
  }
  threshold = sum / 500;
  Serial.print("Initializing pulse oximeter..");

    // Initialize the PulseOximeter instance
  // Failures are generally due to an improper I2C wiring, missing power
    supply
  // or wrong target chip
  if (!pox.begin()){
    Serial.println("FAILED");
    Serial.write(0x07);
    for(;;);
  } else {
    Serial.println("SUCCESS");
  }
  // The default current for the IR LED is 50mA and it could be changed
  // Check MAX30100_Registers.h for all the available options.
```

```
59    pox.setIRLedCurrent(MAX30100_LED_CURR_11MA);
60
61    // Register a callback for the beat detection
62    pox.setOnBeatDetectedCallback(onBeatDetected);
63  }
64
65  void loop(){
66    int gsr;
67    float bpm, spO2;
68    // Make sure to call update as fast as possible
69    pox.update();
70
71      // Asynchronously dump heart rate and oxidation levels to the serial
72    // For both, a value of 0 means "invalid"
73    if (millis() - tsLastReport > REPORTING_PERIOD_MS) {
74      bpm = bpm_aux;
75      spO2 = pox.getSpO2();
76      sensorValueGSR = analogRead(GSR);
77      gsr = threshold - sensorValueGSR;
78
79      imprimeGSR("{GSR;", gsr);
80      Serial.print(";");
81      imprimeOximetro("BPM;", bpm , ";SaO2;", spO2);
82      Serial.println("}");
83      tsLastReport = millis();
84    }
85  }
86
87  void imprimeGSR(const char * leftStr, int MyVar){
88    Serial.print(leftStr);
89    Serial.print(MyVar);
90  }
91
92  void imprimeOximetro(const char *leftStr, float MyVar, const char *sepStr
      , float MyVar1){
93    Serial.print(leftStr);
94    Serial.print(MyVar);
95    Serial.print(sepStr);
96    Serial.print(MyVar1);
97  }
```

Apêndice D

Questionário Pré-Teste

Delineamento do Perfil do Participante

Faixa etária:

( ) 18 a 23 anos ( ) 24 a 29 anos ( ) 30 a 35 anos ( ) Acima de 35 anos

Seu grau de instrução:

( ) Ensino Médio Incompleto ( ) Ensino Médio Completo ( ) Superior Incompleto

( ) Superior Completo ( ) Pós-graduação Incompleta ( ) Pós-graduação Completa

( ) Outro

Você se identifica em qual gênero?

( ) Masculino ( ) Feminino

Entendimento em Línguas Estrangeiras

Você consegue ouvir e entender o que um interlocutor está falando em inglês?

( ) Sim ( ) Não

Você entende frases escritas em inglês?

( ) Sim ( ) Não

Você entende frases escritas em espanhol?

( ) Sim ( ) Não

Acuidade Visual

Você necessita de corretivos visuais (óculos ou lentes)?

( ) Sim ( ) Não

Você foi diagnosticado com ptose (pálpebra caída)?

( ) Sim ( ) Não

Você foi diagnosticado com pterígio ("carne no olho", pequena membrana que cresce sobre a superfície do olho)?

( ) Sim ( ) Não

Você foi diagnosticado com alguma enfermidade ocular?

( ) Sim ( ) Não

Você já realizou alguma cirurgia corretiva?

( ) Sim ( ) Não

Vídeos

Com que freqüência assiste a vídeos digitais (filmes, séries, postagens, etc)?

( ) Várias vezes ao dia ( ) Todos os dias

( ) Mais ou menos três vezes na semana ( ) Menos de três vezes na semana

( ) Não assiste vídeos digitais

Quais são os três conteúdos de vídeo mais freqüentemente assistido por você? (escolha apenas três)

( ) Filmes ( ) Séries

( ) Vídeos de música ( ) Tutoriais

( ) Esportes (eventos na integra) ( ) Conteúdos produzidos por usuário

( ) Notícias / Resumos de esporte ( ) Trailers de filmes

( ) Transmissões ao vivo ( ) Concertos

( ) Outros

Se preencheu "Outros", qual(is) é(são) o(s) vídeo(s) mais freqüente(s) assistido por você?_____
_____
_____

Apêndice E

Roteiro de Atividades do Participante nos Primeiro e Segundo Conjuntos de Ensaios

Neste apêndice, expôs-se o roteiro de atividades executadas pelos participantes do Primeiro e Segundo Conjuntos de Ensaios.

Roteiro de Atividades

Descrição:

"Você irá participar de execução de atividades que necessitará a seleção de quadros representativos de um vídeo. É importante que você lembre-se de que, em nenhum momento, você será avaliado e que você poderá desistir do teste em qualquer ponto da sessão."

Roteiro:

"Você está assistindo a um conjunto de vídeos e você terá que selecionar quadros de cada vídeo que representem o mais próximo possível o vídeo assistido. Porém tem-se uma quantidade máxima de quadros para representar todos os vídeos. Então, estipulou-se que, para cada vídeo, fossem selecionados quadros que correspondessem a ___% do total de quadros do vídeo."

Tempo Total Estimado das Tarefas: 90 minutos.

Observações:

- Sinta-se confortável para perguntar sobre o processo de seleção dos quadros, posterior a exibição do vídeo.
- Durante a exibição do vídeo, mantenha sua cabeça e mãos o mais imóveis possível.

Apêndice F

Roteiro de Atividades do Participante no Terceiro Conjunto de Ensaios

Neste apêndice, apresentou-se o roteiro de atividades executadas pelos participantes do Terceiro Conjunto de Ensaios.

Roteiro de Atividades

Descri´²o:

"Voc¸ ir®participar da execu´²o de atividades que necessitar®a sele´²o de quadros representativos e rotulagem dos quadros de um vºdeo. § importante que voc¸ lembre-se de que, em nenhum momento, voc¸ ser®avaliado e que voc¸ poder®desistir do teste em qualquer ponto da sess²o."

Roteiro:

"Voc¸ est®assistindo a um conjunto de vºdeos e voc¸ ter®que selecionar quadros de cada vºdeo, que representem o mais prÆximo possºvel o vºdeo assistido, por¶m foi determinada uma quantidade mºnima e m®xima de quadros para representar todos os vºdeos e ainda, rotular todos os quadros-chave do vºdeo segundo o sentimento sentido durante sua exibi´²o. Para o processo de selecionar quadros, foi estipulado que, para cada vºdeo, fossem selecionados quadros que correspondessem de 15% a 30% da quantidade total de quadros do vºdeo."

Tempo Total Estimado das Tarefas: 90 minutos.

Observa´Þes:

- ¿ Antes de iniciar o experimento, voc¸ est®confort®vel com o conforto da cadeira e relacionado a temperatura, a ilumina´²o e o som ambiente.
- ¿ Sinta-se confort®vel para perguntar sobre o processo de sele´²o e rotulagem dos quadros, posterior a exibi´²o do vºdeo.
- ¿ Durante a exibi´²o do vºdeo, mantenha sua cabe´a e m²os o mais imÆveis possºvel.

Apêndice G

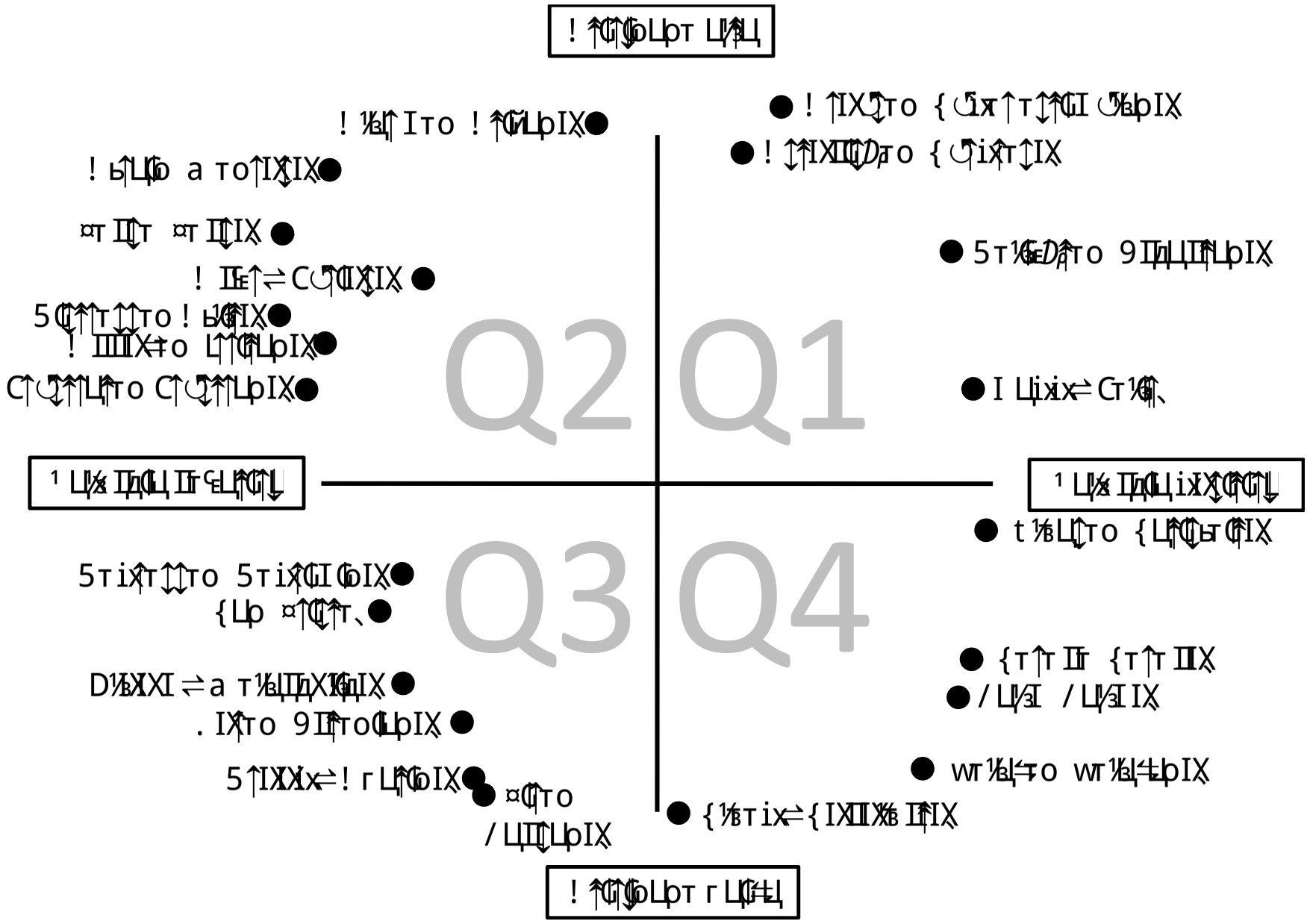
Modelo Circumplexo da Emoção de Russell (1980)

Neste anexo, expõe-se a figura do modelo circumplexo da emoção de Russell, utilizado para rotular os quadros dos vídeos.

A abordagem dimensional denominada Modelo Circumplexo da Emoção de Russell (1980) apresenta os estados afetivos em dois sistemas neurofisiológicos fundamentais, um relacionado à valência (variando de prazer a desagrado) e outro relacionado à excitação ou alerta (variando de excitado a calmo) (CRISPIM et al., 2017). As emoções podem ser expressas como uma combinação dessas duas dimensões. A alegria, por exemplo, é apresentada como um estado emocional que é a combinação de uma ativação forte nos sistemas neurais relacionada à valência, simultaneamente com uma ativação moderada nos sistemas neurais ligadas à excitação.

A dimensão da valência está relacionada a estados emocionais como agradável ou desagradável, enquanto a excitação é a dimensão que corresponde à mobilização ou energia dispensada, sendo representada por um intervalo que varia de baixa ativação (e.g. sono) a alta ativação (e.g. excitação) (RUSSELL; BARRETT, 1999).

Figura G.1: Modelo circumplexo da emoção de Russell (1980)



Fonte: **Russell (1980)**

Apêndice H

Valores Médios do CUS_A Obtidos para Cada Participante em Cada Subconjunto de Características

As Tabelas H.1, H.2 e H.3 contêm os valores médios do CUS_A (apresentados arredondados) obtidos para cada participante em cada subconjunto de características. Nestas tabelas são apresentados também o número de características empregadas em cada treinamento e teste para cada participante em cada subconjunto de características e os valores da CUS_A (mostrados novamente arredondados) obtidos para a seleção aleatória de cada participante. Na última linha das tabelas são apresentadas as médias obtidas para CUS_A e o número médio de características utilizadas para treinamentos e testes em cada subconjuntos de características por participante.

Tabela H.1: Comparação da precisão das estratégias de seleção de características relevantes (M2_5 até M15 e TODAS) e a seleção aleatória por participante

ID	M2_5	M05	M7_5	M10	M12_5	M15	TODAS	Aleat
01	0,274 (20)	0,298 (14)	0,286 (09)	0,286 (05)	0,294 (04)	0,260 (03)	0,280	0,230
02	**0,125** (17)	**0,132** (11)	0,154 (07)	0,175 (03)	**0,146** (02)	0,225 (01)	0,118	0,146
03	0,430 (19)	0,405 (17)	0,432 (14)	0,425 (09)	0,425 (09)	0,428 (08)	0,422	0,228
04	0,256 (22)	0,238 (17)	0,270 (13)	0,274 (10)	0,274 (08)	0,286 (05)	0,248	0,190
05	0,360 (20)	0,374 (19)	0,360 (16)	0,381 (11)	0,364 (06)	0,369 (03)	0,348	0,248
06	0,283 (21)	0,293 (17)	0,302 (15)	0,328 (11)	0,328 (11)	0,307 (09)	0,283	0,202
07	0,252 (22)	0,244 (20)	0,269 (17)	0,277 (14)	0,277 (14)	0,277 (13)	0,250	0,192
08	0,162 (23)	0,164 (19)	0,157 (15)	0,157 (15)	0,155 (13)	0,157 (12)	0,167	0,136
09	0,143 (14)	0,143 (07)	0,187 (05)	0,187 (04)	0,187 (04)	0,210 (03)	**0,117**	0,130
10	0,279 (22)	0,304 (13)	0,331 (05)	0,297 (04)	0,297 (04)	0,303 (03)	0,271	0,228
11	0,266 (18)	0,282 (14)	0,290 (13)	0,278 (11)	0,270 (09)	0,284 (08)	0,254	0,154
12	0,392 (23)	0,392 (22)	0,400 (16)	0,406 (15)	0,421 (11)	0,400 (09)	0,385	0,200
13	0,245 (21)	0,252 (19)	0,239 (15)	0,239 (12)	0,250 (11)	0,234 (10)	0,252	0,152
14	0,215 (20)	0,221 (19)	0,244 (14)	0,260 (10)	0,252 (09)	0,275 (07)	0,210	0,198
15	0,190 (21)	0,180 (17)	0,212 (14)	0,222 (09)	0,202 (05)	0,202 (05)	0,190	0,172
16	0,200 (17)	0,200 (10)	0,192 (07)	0,197 (06)	0,214 (05)	0,197 (04)	0,203	0,175
17	0,382 (19)	0,373 (16)	0,392 (12)	0,392 (09)	0,392 (09)	0,392 (07)	0,373	0,230
18	0,188 (18)	0,175 (13)	0,228 (08)	0,252 (05)	0,360 (02)	0,360 (02)	0,202	0,155
19	0,393 (22)	0,393 (22)	0,393 (22)	0,393 (22)	0,409 (20)	0,400 (18)	0,403	0,221
20	0,233 (19)	0,226 (16)	0,264 (12)	0,288 (10)	0,267 (09)	0,267 (09)	0,231	0,148
21	0,428 (19)	0,444 (16)	0,443 (12)	0,450 (10)	0,437 (07)	0,437 (07)	0,398	0,204
22	0,410 (24)	0,426 (19)	0,439 (17)	0,439 (13)	0,439 (11)	0,461 (09)	0,426	0,239
23	0,296 (18)	0,301 (14)	0,305 (13)	0,322 (07)	0,318 (06)	0,311 (03)	0,314	0,268
24	0,227 (22)	0,225 (21)	0,279 (15)	0,258 (09)	0,242 (05)	0,233 (03)	0,227	0,183
25	0,218 (22)	0,240 (16)	0,240 (16)	0,265 (12)	0,265 (12)	0,240 (11)	0,220	0,142
26	0,311 (22)	0,328 (18)	0,317 (15)	0,311 (12)	0,344 (09)	0,347 (08)	0,283	0,147
27	0,265 (22)	0,250 (20)	0,281 (16)	0,271 (13)	0,262 (08)	0,262 (08)	0,277	0,169
28	0,273 (20)	0,295 (18)	0,293 (15)	0,302 (12)	0,300 (11)	0,300 (11)	0,275	0,155
29	0,252 (21)	0,275 (12)	0,306 (08)	0,296 (07)	0,310 (06)	0,327 (05)	0,254	0,200
30	0,308 (21)	0,331 (20)	0,344 (16)	0,338 (15)	0,338 (15)	0,346 (13)	0,302	0,148
31	0,184 (21)	0,184 (18)	0,191 (17)	0,197 (16)	0,209 (15)	0,175 (14)	0,181	0,125
32	0,278 (21)	0,312 (15)	0,336 (12)	0,348 (11)	0,358 (09)	0,331 (06)	0,266	0,234
33	0,255 (21)	0,239 (14)	0,252 (13)	0,261 (09)	0,241 (07)	0,239 (03)	0,257	0,200
Media	0,272 (20,4)	0,277 (16,5)	0,292 (13,2)	0,296 (10,3)	0,298 (8,7)	0,298 (7,3)	0,269	0,186

Fonte: **Autor**.

Nota: As colunas contêm o ID do participante, as estratégias de características relevantes e a seleção aleatória, todas com CUS_A e o número de características usadas no subconjunto. Foi utilizada a máquina de aprendizagem *Random Forest* no treinamento e teste. Os valores em negrito indicam os casos em que uma estratégia teve desempenho pior ou igual à seleção aleatória.

Tabela H.2: Comparação da precisão das estratégias de seleção de características relevantes (S30 até S55 e TODAS) e a seleção aleatória por participante

ID	S30	S32_5	S35	S37_5	S40	S42_5	S45	S47_5	S50	S52_5	S55	TODAS	Aleat
01	0,268 (02)	0,268 (02)	0,268 (02)	0,268 (02)	0,260 (03)	0,260 (03)	0,260 (03)	0,294 (04)	0,294 (04)	0,294 (04)	0,286 (05)	0,280	0,230
02	**0,146** (02)	**0,146** (02)	**0,146** (02)	**0,146** (02)	0,175 (03)	0,175 (03)	0,175 (03)	**0,132** (04)	**0,132** (04)	0,150 (05)	0,150 (05)	**0,118**	0,146
03	0,372 (01)	0,372 (01)	0,375 (02)	0,375 (02)	0,375 (02)	0,375 (02)	0,348 (03)	0,348 (03)	0,348 (03)	0,392 (04)	0,392 (04)	0,422	0,228
04	0,246 (02)	0,264 (03)	0,264 (03)	0,264 (03)	0,264 (03)	0,260 (04)	0,260 (04)	0,260 (04)	0,286 (05)	0,286 (05)	0,270 (06)	0,248	0,190
05	0,369 (03)	0,369 (03)	0,350 (04)	0,350 (04)	0,350 (04)	0,357 (05)	0,357 (05)	0,364 (06)	0,364 (06)	0,350 (07)	0,350 (07)	0,348	0,248
06	0,302 (02)	0,302 (02)	0,302 (02)	0,327 (03)	0,327 (03)	0,327 (03)	0,327 (03)	0,315 (04)	0,315 (04)	0,362 (05)	0,362 (05)	0,283	0,202
07	0,310 (03)	0,310 (03)	0,310 (03)	0,296 (04)	0,296 (04)	0,296 (04)	0,292 (05)	0,292 (05)	0,292 (05)	0,304 (06)	0,304 (06)	0,250	0,192
08	0,167 (03)	0,169 (04)	0,169 (04)	0,169 (04)	0,171 (05)	0,171 (05)	0,171 (05)	0,179 (06)	0,179 (06)	0,179 (06)	0,174 (07)	0,167	0,136
09	0,230 (01)	0,230 (01)	0,230 (01)	0,230 (01)	0,230 (01)	0,200 (02)	0,200 (02)	0,200 (02)	0,200 (02)	0,200 (02)	0,200 (02)	**0,117**	0,130
10	0,368 (01)	0,368 (01)	0,368 (01)	0,368 (01)	0,322 (02)	0,322 (02)	0,322 (02)	0,322 (02)	0,303 (03)	0,303 (03)	0,303 (03)	0,271	0,228
11	0,270 (01)	0,246 (02)	0,246 (02)	0,246 (02)	0,246 (02)	0,246 (02)	0,246 (02)	0,256 (03)	0,256 (03)	0,244 (04)	0,244 (04)	0,254	0,154
12	0,435 (02)	0,435 (02)	0,435 (02)	0,446 (03)	0,446 (03)	0,446 (03)	0,446 (03)	0,406 (04)	0,406 (04)	0,406 (04)	0,406 (04)	0,385	0,200
13	0,227 (02)	0,191 (03)	0,191 (03)	0,191 (03)	0,216 (04)	0,216 (04)	0,216 (04)	0,205 (05)	0,205 (05)	0,205 (05)	0,218 (06)	0,252	0,152
14	0,283 (03)	0,283 (03)	0,283 (03)	0,283 (03)	0,271 (04)	0,271 (04)	0,269 (05)	0,269 (05)	0,269 (05)	0,281 (06)	0,281 (06)	0,210	0,198
15	0,188 (03)	0,188 (03)	0,188 (03)	0,200 (04)	0,200 (04)	0,200 (04)	0,202 (05)	0,202 (05)	0,202 (05)	0,242 (06)	0,242 (06)	0,190	0,172
16	0,231 (01)	0,231 (01)	0,211 (02)	0,211 (02)	0,211 (02)	0,211 (02)	0,211 (02)	0,200 (03)	0,200 (03)	0,200 (03)	0,200 (03)	0,203	0,175
17	0,400 (02)	0,400 (02)	0,400 (02)	0,400 (02)	0,400 (02)	0,400 (02)	0,425 (03)	0,425 (03)	0,425 (03)	0,425 (03)	0,425 (03)	0,373	0,230
18	0,352 (01)	0,360 (02)	0,360 (02)	0,360 (02)	0,255 (03)	0,255 (03)	0,255 (03)	0,252 (04)	0,252 (04)	0,252 (04)	0,252 (05)	0,202	0,155
19	0,353 (03)	0,369 (04)	0,369 (04)	0,369 (04)	0,383 (05)	0,383 (05)	0,390 (06)	0,390 (06)	0,416 (07)	0,416 (07)	0,411 (08)	0,403	0,221
20	0,295 (02)	0,295 (02)	0,295 (02)	0,262 (03)	0,262 (03)	0,262 (03)	0,262 (03)	0,252 (04)	0,252 (04)	0,252 (04)	0,262 (05)	0,231	0,148
21	0,276 (01)	0,276 (01)	0,276 (01)	0,276 (01)	0,378 (02)	0,378 (02)	0,378 (02)	0,378 (02)	0,378 (02)	0,396 (03)	0,396 (03)	0,398	0,204
22	0,381 (02)	0,381 (02)	0,381 (02)	0,381 (02)	0,445 (03)	0,445 (03)	0,445 (03)	0,458 (04)	0,458 (04)	0,458 (04)	0,453 (05)	0,426	0,239
23	0,426 (01)	0,426 (01)	**0,259** (02)	**0,259** (02)	**0,259** (02)	0,311 (03)	0,311 (03)	0,311 (03)	0,301 (04)	0,301 (04)	0,323 (05)	0,314	0,268
24	0,842 (01)	0,842 (01)	0,248 (02)	0,248 (02)	0,248 (02)	0,233 (03)	0,233 (03)	0,212 (04)	0,212 (04)	0,242 (05)	0,246 (06)	0,227	0,183
25	0,398 (01)	0,312 (02)	0,312 (02)	0,312 (02)	0,312 (02)	0,310 (03)	0,310 (03)	0,310 (03)	0,280 (04)	0,280 (04)	0,280 (04)	0,220	0,142
26	0,306 (03)	0,306 (03)	0,306 (03)	0,356 (04)	0,356 (04)	0,372 (05)	0,372 (05)	0,372 (05)	0,367 (06)	0,367 (06)	0,367 (06)	0,283	0,147
27	0,269 (02)	0,269 (02)	0,269 (02)	0,273 (03)	0,273 (03)	0,273 (03)	0,273 (03)	0,265 (04)	0,265 (04)	0,265 (04)	0,265 (05)	0,277	0,169
28	0,284 (02)	0,284 (02)	0,257 (03)	0,257 (03)	0,257 (03)	0,257 (03)	0,275 (04)	0,275 (04)	0,286 (05)	0,286 (05)	0,286 (05)	0,275	0,155
29	0,283 (01)	0,242 (02)	0,242 (02)	0,242 (02)	0,242 (02)	0,242 (02)	0,315 (03)	0,315 (03)	0,315 (03)	0,315 (03)	0,312 (04)	0,254	0,200
30	0,238 (01)	0,265 (02)	0,265 (02)	0,265 (02)	0,265 (02)	0,283 (03)	0,283 (03)	0,279 (04)	0,279 (04)	0,279 (04)	0,310 (05)	0,302	0,148
31	0,181 (02)	0,181 (02)	0,166 (03)	0,166 (03)	0,166 (03)	0,166 (03)	0,144 (04)	0,144 (04)	0,144 (04)	0,150 (05)	0,150 (05)	0,181	0,125
32	0,330 (01)	0,353 (02)	0,353 (02)	0,353 (02)	0,353 (02)	0,333 (03)	0,333 (03)	0,333 (03)	0,330 (04)	0,330 (04)	0,350 (05)	0,266	0,234
33	0,239 (03)	0,239 (03)	0,239 (03)	0,239 (04)	0,239 (04)	0,239 (04)	0,255 (05)	0,255 (05)	0,255 (05)	0,230 (06)	0,230 (06)	0,257	0,200
Media	0,311 (1,8)	0,308 (2,2)	0,283 (2,4)	0,284 (2,6)	0,286 (2,9)	0,287 (3,2)	0,290 (3,5)	0,287 (3,9)	0,287 (4,2)	0,292 (4,5)	0,294 (5)	0,269	0,186

Fonte: **Autor**.

Nota: As colunas contêm o ID do participante, as estratégias de características relevantes e a seleção aleatória, todas com CUS_A e o número de características usadas no subconjunto. Foi utilizada a máquina de aprendizagem *Random Forest* no treinamento e teste. Os valores em negrito indicam os casos em que uma estratégia teve desempenho pior ou igual à seleção aleatória.

Tabela H.3: Comparação da precisão das estratégias de seleção de características relevantes (S57_5 até S80 e TODAS) e a seleção aleatória por participante

ID	S57_5	S60	S62_5	S65	S67_5	S70	S72_5	S75	S77_5	S80	TODAS	Aleat
01	0,286 (05)	0,288 (06)	0,288 (06)	0,290 (07)	0,290 (07)	0,276 (08)	0,276 (08)	0,286 (09)	0,294 (10)	0,292 (11)	0,280	0,230
02	0,150 (05)	0,171 (06)	0,171 (06)	0,171 (06)	0,154 (07)	0,154 (07)	0,186 (08)	0,186 (08)	0,186 (09)	0,186 (09)	**0,118**	0,146
03	0,392 (04)	0,392 (04)	0,432 (05)	0,432 (05)	0,432 (05)	0,418 (06)	0,418 (06)	0,428 (07)	0,428 (07)	0,428 (08)	0,422	0,228
04	0,270 (06)	0,276 (07)	0,276 (07)	0,276 (07)	0,274 (08)	0,274 (08)	0,270 (09)	0,274 (10)	0,274 (10)	0,270 (11)	0,248	0,190
05	0,360 (08)	0,360 (08)	0,352 (09)	0,352 (09)	0,390 (10)	0,390 (10)	0,381 (11)	0,381 (11)	0,371 (12)	0,374 (13)	0,348	0,248
06	0,340 (06)	0,340 (06)	0,323 (07)	0,323 (07)	0,323 (07)	0,308 (08)	0,308 (08)	0,307 (09)	0,327 (10)	0,327 (10)	0,283	0,202
07	0,265 (07)	0,265 (07)	0,300 (08)	0,300 (08)	0,290 (09)	0,290 (09)	0,277 (10)	0,277 (10)	0,271 (11)	0,269 (12)	0,250	0,192
08	0,174 (07)	0,169 (08)	0,169 (08)	0,169 (08)	0,171 (09)	0,171 (09)	0,176 (10)	0,176 (10)	0,164 (11)	0,164 (11)	0,167	0,136
09	0,210 (03)	0,210 (03)	0,210 (03)	0,210 (03)	0,187 (04)	0,187 (04)	0,187 (04)	0,187 (05)	0,187 (05)	0,170 (06)	0,117	0,130
10	0,297 (04)	0,297 (04)	0,331 (05)	0,319 (06)	0,315 (07)	0,315 (07)	0,328 (08)	0,321 (09)	0,319 (10)	0,310 (11)	0,271	0,228
11	0,244 (04)	0,250 (05)	0,250 (05)	0,258 (06)	0,258 (06)	0,258 (06)	0,264 (07)	0,264 (07)	0,284 (08)	0,284 (08)	0,254	0,154
12	0,419 (05)	0,419 (05)	0,419 (05)	0,440 (06)	0,440 (06)	0,425 (07)	0,425 (07)	0,419 (08)	0,419 (08)	0,400 (09)	0,385	0,200
13	0,218 (06)	0,218 (06)	0,191 (07)	0,191 (07)	0,191 (07)	0,211 (08)	0,211 (08)	0,223 (09)	0,223 (09)	0,234 (10)	0,252	0,152
14	0,281 (06)	0,275 (07)	0,275 (07)	0,277 (08)	0,277 (08)	0,252 (09)	0,252 (09)	0,260 (10)	0,260 (10)	0,250 (11)	0,210	0,198
15	0,245 (07)	0,245 (07)	0,222 (08)	0,222 (08)	0,222 (09)	0,222 (09)	0,212 (10)	0,212 (11)	0,212 (11)	0,208 (12)	0,190	0,172
16	0,200 (03)	0,197 (04)	0,197 (04)	0,197 (04)	0,214 (05)	0,214 (05)	0,214 (05)	0,197 (06)	0,197 (06)	0,192 (07)	0,203	0,175
17	0,422 (04)	0,422 (04)	0,422 (04)	0,422 (05)	0,422 (05)	0,422 (05)	0,408 (06)	0,408 (06)	0,392 (07)	0,392 (07)	0,373	0,230
18	0,252 (05)	0,252 (05)	0,250 (06)	0,250 (06)	0,250 (07)	0,250 (07)	0,228 (08)	0,228 (08)	0,210 (09)	0,212 (10)	0,202	0,155
19	0,411 (08)	0,419 (09)	0,419 (09)	0,416 (10)	0,416 (10)	0,420 (11)	0,420 (11)	0,409 (12)	0,401 (13)	0,401 (13)	0,403	0,221
20	0,262 (05)	0,276 (06)	0,276 (06)	0,276 (06)	0,271 (07)	0,271 (07)	0,243 (08)	0,243 (08)	0,243 (08)	0,267 (09)	0,231	0,148
21	0,396 (03)	0,417 (04)	0,417 (04)	0,417 (04)	0,443 (05)	0,443 (05)	0,443 (05)	0,439 (06)	0,439 (06)	0,437 (07)	0,398	0,204
22	0,453 (05)	0,453 (05)	0,445 (06)	0,445 (06)	0,445 (06)	0,442 (07)	0,442 (07)	0,450 (08)	0,450 (08)	0,461 (09)	0,426	0,239
23	0,323 (05)	0,323 (05)	0,318 (06)	0,318 (06)	0,322 (07)	0,322 (07)	0,301 (08)	0,301 (08)	0,323 (09)	0,319 (10)	0,314	0,268
24	0,246 (06)	0,235 (07)	0,235 (07)	0,265 (08)	0,258 (09)	0,258 (09)	0,252 (10)	0,248 (11)	0,252 (12)	0,263 (13)	0,227	0,183
25	0,245 (05)	0,245 (05)	0,248 (06)	0,248 (06)	0,248 (06)	0,248 (07)	0,248 (07)	0,265 (08)	0,245 (09)	0,245 (09)	0,220	0,142
26	0,367 (07)	0,367 (07)	0,367 (07)	0,347 (08)	0,347 (08)	0,344 (09)	0,344 (09)	0,317 (10)	0,319 (11)	0,319 (11)	0,283	0,147
27	0,265 (05)	0,265 (05)	0,267 (06)	0,267 (06)	0,254 (07)	0,262 (08)	0,262 (08)	0,273 (09)	0,279 (10)	0,290 (11)	0,277	0,169
28	0,295 (06)	0,295 (06)	0,295 (06)	0,298 (07)	0,298 (07)	0,311 (08)	0,311 (08)	0,298 (09)	0,298 (09)	0,298 (09)	0,275	0,155
29	0,312 (04)	0,312 (04)	0,312 (04)	0,327 (05)	0,327 (05)	0,310 (06)	0,310 (06)	0,310 (06)	0,296 (07)	0,306 (08)	0,254	0,200
30	0,310 (05)	0,333 (06)	0,333 (06)	0,333 (06)	0,371 (07)	0,350 (08)	0,350 (08)	0,346 (09)	0,346 (09)	0,346 (10)	0,302	0,148
31	0,150 (05)	0,178 (06)	0,178 (06)	0,178 (06)	0,144 (07)	0,172 (08)	0,172 (08)	0,153 (09)	0,166 (10)	0,156 (11)	0,181	0,125
32	0,350 (05)	0,350 (05)	0,331 (06)	0,331 (06)	0,372 (07)	0,372 (07)	0,364 (08)	0,364 (08)	0,358 (09)	0,358 (09)	0,266	0,234
33	0,230 (06)	0,241 (07)	0,241 (07)	0,266 (08)	0,266 (08)	0,261 (09)	0,261 (09)	0,239 (10)	0,239 (10)	0,239 (11)	0,257	0,200
Media	0,292 (5,3)	0,296 (5,7)	0,296 (6,1)	0,298 (6,5)	0,299 (7,0)	0,298 (7,5)	0,295 (7,9)	0,294 (8,6)	0,293 (9,2)	0,293 (9,9)	0,269	0,186

Fonte: **Autor**.

Nota: As colunas contêm o ID do participante, as estratégias de características relevantes e a seleção aleatória, todas com CUS_A e o número de características usadas no subconjunto. Foi utilizada a máquina de aprendizagem *Random Forest* no treinamento e teste. Os valores em negrito indicam os casos em que uma estratégia teve desempenho pior ou igual à seleção aleatória.

Apêndice I

Características Mais Relevantes por Subconjunto de Características

Nas Tabelas I.1 e I.2, são apresentadas as características mais representativas por subconjunto de características. Nas tabelas é contabilizado o número de participantes em que cada característica foi relevante em cada subconjunto.

Tabela I.1: Características consideradas relevantes pelo seletor Boruta nos subconjuntos M2_5 até M15, apresentando-se o número de participantes

Caract.	**M2_5**	M05	M7_5	**M10**	M12_5	M15
GSR	29	28	21	19	15	11
BPM	29	26	19	12	9	6
AU01	30	24	22	16	12	11
AU02	31	25	22	19	19	17
AU04	29	24	18	14	12	10
AU05	23	16	15	9	7	7
AU06	32	31	27	22	17	15
AU07	28	25	21	18	16	15
AU09	30	28	20	18	14	14
AU10	29	25	20	17	15	14
AU12	30	27	24	22	19	16
AU14	30	30	26	19	18	15
AU15	31	24	17	14	10	7
AU17	31	24	23	16	16	12
AU20	28	24	20	15	14	12
AU23	23	15	13	10	8	6
AU25	32	24	20	18	13	12
AU26	30	26	21	15	11	8
AU45	31	23	17	14	12	9
Fix0_5s	20	12	5	3	3	2
Fix1s	23	13	6	5	3	3
Fix1_5s	5	2	1	0	0	0
Fix2s	2	2	2	1	1	0
Fix	19	14	10	8	6	4
NFix	24	15	11	5	5	4
Quadrant	23	16	13	12	11	10

Fonte: **Autor**.
Nota: Colunas realçadas representam os subconjuntos que obtiveram o pior resultado e aquela que obteve o melhor resultado na análise quando comparado à seleção aleatória.

Tabela I.2: Características consideradas relevantes pelo seletor Boruta nos subconjuntos S30 até S80, apresentando-se o número de participantes

Caract.	S30	S32_5	S35	S37_5	S40	S42_5	S45	S47_5	S50	S52_5	S55	S57_5	S60	S62_5	**S65**	S67_5	S70	S72_5	S75	S77_5	S80
GSR	2	2	2	2	2	2	4	4	6	7	7	8	9	10	11	11	13	14	16	16	16
BPM	3	3	3	3	3	3	4	5	5	6	6	6	6	6	6	6	6	6	6	7	8
AU01	2	2	2	3	4	4	4	4	4	4	5	5	6	6	6	9	9	9	10	11	11
AU02	3	4	4	4	5	5	5	5	5	5	7	7	8	8	11	12	13	14	14	15	17
AU04	1	1	1	1	1	1	2	2	2	2	3	3	3	4	4	6	6	7	9	10	12
AU05	2	2	2	2	2	2	2	2	2	2	2	2	3	5	5	6	8	8	8	8	8
AU06	3	5	6	6	6	8	8	10	10	10	10	11	13	14	14	15	17	18	19	21	23
AU07	2	4	4	4	6	6	6	7	9	11	11	12	12	12	12	12	12	12	14	16	18
AU09	4	5	6	6	7	7	8	9	9	10	10	10	11	12	13	13	13	13	14	14	15
AU10	4	5	5	5	5	5	5	6	7	7	7	10	11	13	13	14	15	16	16	18	18
AU12	5	5	5	5	6	6	6	8	9	9	9	9	11	12	15	16	16	17	18	18	20
AU14	3	3	3	6	6	8	9	10	10	13	14	14	15	15	15	15	16	17	18	19	22
AU15	2	2	2	2	2	3	5	6	6	6	8	9	10	10	10	11	11	11	13	13	15
AU17	2	4	5	5	5	5	5	7	7	7	8	9	9	9	9	11	12	14	16	17	17
AU20	4	4	4	5	5	5	5	5	6	8	8	9	10	10	11	12	12	12	12	14	16
AU23	1	2	2	2	2	2	3	3	3	3	4	5	5	5	5	5	6	7	8	8	9
AU25	5	5	6	7	7	7	7	7	7	7	10	10	10	11	11	12	14	15	16	16	16
AU26	2	2	3	3	3	4	5	5	6	6	6	6	6	6	8	10	12	12	12	12	13
AU45	1	1	3	3	5	7	7	7	7	8	8	8	8	8	8	8	8	8	10	11	11
Fix0_5s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	4	4	6	7
Fix1s	2	2	2	2	2	2	2	3	3	3	3	3	4	4	5	5	5	5	6	7	7
Fix1_5s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fix2s	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
Fix	1	1	1	1	2	3	3	3	3	4	4	4	4	6	6	6	6	6	7	7	8
NFix	2	2	2	3	3	3	3	4	4	4	4	4	4	4	4	4	4	5	6	7	7
Quadrant	5	5	5	6	7	7	7	8	8	8	10	11	11	11	11	11	11	11	11	11	11

Fonte: **Autor**.
Nota: Coluna realçada representa o subconjunto que obteve o melhor resultado na análise quando comparado à seleção aleatória.

Anexo I

Documentação

Neste anexo, apresentamos os documentos referentes ao Comitê de Ética em Pesquisas (CEP) com Seres Humanos.

I.1 Documentação Requisitada para Protocolar o Projeto Junto ao CEP

- Declaração de Divulgação dos Resultados ;
- Termo de Compromisso do Pesquisador ;
- Termo de Consentimento Livre e Esclarecido ; e
- Termo de Anuência Institucional .

Termo de Compromisso de divulgação dos resultados

Por este termo de responsabilidade, nós, abaixo – assinados, respectivamente, autor e orientando da pesquisa intitulada "**SUMARIZAÇÃO DE VÍDEOS À PARTIR DE DADOS FISIOLÓGICOS E EXPRESSÕES FACIAIS**" assumimos o compromisso de:

- Preservar a privacidade dos participantes da pesquisa cujos dados serão coletados;
- Assegurar que as informações serão utilizadas única e exclusivamente para a execução do projeto em questão;
- Assegurar que os benefícios resultantes do projeto retornem aos participantes da pesquisa, seja em termos de retorno social, acesso aos procedimentos, produtos ou agentes da pesquisa;
- Assegurar que as informações somente serão divulgadas de forma anônima, não sendo usadas iniciais ou quaisquer outras indicações que possam identificar o sujeito da pesquisa;
- Assegurar que os resultados da pesquisa serão encaminhados para a publicação, com os devidos créditos aos autores.

Campina Grande, 12 de abril de 2018.

Herman Martins Gomes
Orientador

Sérgio Cavalcanti de Paiva
Sérgio Cavalcanti de Paiva
Orientando

Termo de Compromisso do (s) Pesquisador (es)

Por este termo de responsabilidade, nós, abaixo – assinados, respectivamente, autor e orientando da pesquisa intitulada "**Sumarização de Vídeos à partir de Dados Fisiológicos e Expressões Faciais**" assumimos cumprir fielmente as diretrizes regulamentadoras emanadas da Resolução nº 466, de 12 deDezembro de 2012do Conselho Nacional de Saúde/ MS e suas Complementares, homologada nos termos do Decreto de Delegação de Competência de 12 de novembro de 1991, visando assegurar os direitos e deveres que dizem respeito à comunidade cientifica, ao (s) sujeito (s) da pesquisa e ao Estado.

Reafirmamos, outrossim, nossa responsabilidade indelegável e intransferível, mantendo em arquivo todas as informações inerentes a presente pesquisa, respeitando a confidencialidade e sigilo das fichas correspondentes a cada sujeito incluído na pesquisa, por um período de 5 (cinco) anos após o término desta. Apresentaremos sempre que solicitado pelo CEP/ HUAC (Comitê de Ética em Pesquisas/ Hospital Universitário Alcides Carneiro), ou CONEP (Comissão Nacional de Ética em Pesquisa) ou, ainda, as Curadorias envolvidas no presente estudo, relatório sobre o andamento da pesquisa, comunicando ainda ao CEP/ HUAC, qualquer eventual modificação proposta no supracitado projeto.

Campina Grande, 19 de março de 2018.

Herman Martins Gomes
Orientador

Sérgio Cavalcanti de Paiva
Orientando

UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
HOSPITAL UNIVERSITÁRIO ALCIDES CARNEIRO
Comitê de ÉticaemPesquisas com SeresHumanos - CEP
Rua: Dr. Carlos Chagas, s/ n, São José. CEP: 58107 – 670.
Tel: 2101 – 5545, E-mail: cep@huac.ufcg.edu.br.

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Sumarização de Vídeos à Partir de Dados Fisiológicos e Expressões Faciais

Você está sendo convidado (a) a participar do projeto de pesquisa acima citado. O documento abaixo contém todas as informações necessárias sobre a pesquisa que estamos fazendo. Sua colaboração neste estudo será de muita importância para nós, mas se desistir a qualquer momento, isso não causará nenhum prejuízo a você.

Eu, _____, profissão_____. residente e domiciliado na _____, portador da Cédula de identidade, RG _____ e inscrito no CPF _____, nascido(a) em ___ / _____ /_____, abaixo assinado(a), concordo de livre e espontânea vontade em participar como voluntário(a) do estudo "Sumarização de Vídeos à Partir de Dados Fisiológicos e Expressões Faciais". Declaro que obtive todas as informações necessárias, bem como a promessa dos esclarecimentos às dúvidas, por mim apresentadas durante o decorrer da pesquisa.

Estou ciente que:

I) A participação neste projeto não tem objetivo de me submeter a um tratamento., bem como não me acarretará qualquer ônus pecuniário. Será garantido a indenização diante de eventuais danos decorrentes da pesquisa.

II) Tenho a liberdade de desistir ou de interromper a colaboração neste estudo no momento em que desejar, sem necessidade de qualquer explicação;

III) Os resultados obtidos durante este ensaio serão mantidos em sigilo, mas concordo que sejam divulgados em publicações científicas, desde que meus dados pessoais não sejam mencionados;

IV) Caso deseje, poderei pessoalmente tomar conhecimento dos resultados, ao final desta pesquisa. Estou ciente que receberei uma via deste termo de consentimento;

( ) Desejo conhecer os resultados desta pesquisa.

( ) Não desejo conhecer os resultados desta pesquisa.

V) Esta pesquisa é de caráter voluntário, portanto não haverá compensação financeira ou custos decorrentes de minha participação. Poderei retirar todas as dúvidas, durante e após o estudo, havendo o compromisso do pesquisador em respondê-las.

VI) A coleta de dados será composta de vídeos e registros fisiológicos, porém, é garantido o sigilo e a confidencialidade dos mesmos, sendo divulgados apenas em eventos e publicações científicas, preservando sempre minha identidade.

VII) Partindo do pressuposto de que toda pesquisa com seres humanos envolve riscos de caráter e dimensões variantes, esta pesquisa segue a Resolução nº 466/12 (BRASIL, 2012), ponderando alguns pontos, tanto para os pesquisadores, como para os pesquisados.

UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
HOSPITAL UNIVERSITÁRIO ALCIDES CARNEIRO
Comitê de ÉticaemPesquisas com SeresHumanos - CEP
Rua: Dr. Carlos Chagas, s/ n, São José. CEP: 58107 – 670.
Tel: 2101 – 5545, E-mail: cep@huac.ufcg.edu.br.

Riscos

- Partindo do pressuposto de que toda pesquisa com seres humanos envolve riscos de caráter e dimensões variantes, esta pesquisa segue a Resolução no 466/12 CNS/MS (BRASIL. Conselho Nacional de Saúde, 2012), ponderando alguns pontos, tanto para os pesquisadores, como para os pesquisados.
- Possíveis desconfortos, constrangimentos e impertinência durante a aplicação do objeto de coleta de dados pode acontecer, devido à ocupação do tempo para a realização do experimento. Incluindo ainda a não aceitação, em participar da pesquisa e a possibilidade da utilização de informações, pertinentes à pesquisa, por terceiros e que possam influenciar o vínculo dos objetos de estudo com a comunidade.
- Acentua-se que os participantes terão a possibilidade de desistir de participar da pesquisa a qualquer momento, que lhes serão garantidos o sigilo dos discursos, com o uso das siglas para cada participante da pesquisa e que o experimento realizar-se-á após consentimento dos (as) pesquisados (as).

Benefícios

- Podemos destacar como benefícios para a sociedade, a concepção de um método capaz de sintetizar vídeos digitais de maneira efetiva e compreensiva, levando em conta preferências individuais em uma área em que os telespectadores estão geralmente sobrecarregados com o volume de informação ao seu dispor.
- Para a comunidade acadêmica a ampliação dos conhecimentos científicos acerca da temática e possibilidade de novas pesquisas.
- Os pesquisadores se comprometem a atender a todas as exigências necessárias à pesquisa com seres humanos conforme preconiza, a Resolução 466/2012 CNS (Termo de Compromisso do Pesquisador - Anexo 02).

Como minimizar os riscos expostos no TCLE

Garantir o acesso aos resultados individuais e coletivos; Minimizar desconfortos, garantindo local reservado; Estar atento aos sinais verbais e não verbais de desconforto; Assegurar a confidencialidade e a privacidade, a proteção da imagem e a não estigmatização, garantindo a não utilização das informações em prejuízo das pessoas e/ou das comunidades, inclusive em termos de autoestima, de prestígio e/ou econômico – financeiro; Assumir a responsabilidade de dar assistência integral às complicações e danos decorrentes dos riscos previstos; Não permitir duplo padrão; Garantir que o estudo será suspenso imediatamente ao perceber algum risco ou dano à saúde do sujeito participante da pesquisa, conseqüentemente à mesma, não previsto no termo de consentimento; Garantir que os sujeitos da pesquisa que vierem a sofrer qualquer tipo de dano previsto ou não no termo de consentimento e resultante de sua participação, além do direito à assistência integral, têm direito à indenização; Garantir a divulgação pública dos resultados; Garantir que sempre serão respeitados os valores culturais, sociais, morais, religiosos e éticos,

UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
HOSPITAL UNIVERSITÁRIO ALCIDES CARNEIRO
Comitê de ÉticaemPesquisas com SeresHumanos - CEP
Rua: Dr. Carlos Chagas, s/ n, São José. CEP: 58107 – 670.
Tel: 2101 – 5545, E-mail: cep@huac.ufcg.edu.br.

bem como os hábitos e costumes quando as pesquisas envolverem comunidades; Assegurar a inexistência de conflito de interesses entre o pesquisador e os sujeitos da pesquisa ou patrocinador do projeto.

O pesquisador se compromete a atender a todas as exigências necessárias à pesquisa com seres humanos conforme preconiza, a Resolução 466/2012 CNS (Termo de Compromisso do Pesquisador

VIII) Caso me sinta prejudicado (a) por participar desta pesquisa, poderei recorrer ao Comitê de Ética em Pesquisas com Seres Humanos – CEP, do Hospital Universitário Alcides Carneiro - HUAC, situado a Rua: Dr. Carlos Chagas, s/ n, São José, CEP: 58401 – 490, Campina Grande-PB, Tel: 2101 – 5545, E-mail: cep@huac.ufcg.edu.br; ao Conselho Regional de Medicina da Paraíba e àDelegacia Regional de Campina Grande.

Campina Grande - PB, _____ de _____de 2017.

( ) Entrevistado / ( ) Responsável: _____.

Testemunha 1 :_____.
Nome / RG / Telefone

Testemunha 2 : _____.
Nome / RG / Telefone

Responsável pelo Projeto: Sérgio Cavalcanti de Paiva.
Mestre em Meteorologia.
Bacharel em Ciência da Computação
Telefone para contato: (83) 999690928

Endereço profissional: Rua Aprígio Veloso, n: 882. Bairro Universitário, Campina Grande – PB.
CEP: 58429-900

TERMO DE ANUÊNCIA INSTITUCIONAL

Eu, Jorge César Abrantes de Figueiredo, diretor do Centro de Engenharia Elétrica e Informática (CEEI) – UFCG, autorizo o desenvolvimento da pesquisa intitulada: **SUMARIZAÇÃO DE VÍDEOS À PARTIR DE DADOS FISIOLÓGICOS E EXPRESSÕES FACIAIS**, neste Programa de Pós Graduação, tendo como pesquisador responsável o Professor Mestre Sérgio Cavalcanti de Paiva.

Campina Grande, 28 de março de 2018.

Jorge César Abrantes de Figueiredo
Diretor do Centro de Engenharia Elétrica e Informática (CEEI) – UFCG

I.2 Documentos da Aprovação pelo CEP

- Parecer Consubstanciado do CEP ; e
- Declaração de Aprovação de Projeto .

UFCG - HOSPITAL
UNIVERSITÁRIO ALCIDES
CARNEIRO DA UNIVERSIDADE

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Sumarização de Vídeos à partir de Dados Fisiológicos e Expressões Faciais
Pesquisador: SERGIO CAVALCANTI DE PAIVA
Área Temática:
Versão: 1
CAAE: 87510318.5.0000.5182
Instituição Proponente: UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 2.618.913

Apresentação do Projeto:
Considerando o aumento de acesso aos vídeos decorrente da revolução tecnológica dos últimos anos, a proposta aborda um tema relevante que poderá trazer importantes benefícios social

Objetivo da Pesquisa:
Desenvolver uma abordagem de sumarização de vídeos relacionada com o feedback fisiológico e expressões faciais do telespectador durante a exibição do vídeo

Avaliação dos Riscos e Benefícios:
O proponente prever situações de riscos como desconforto e constrangimento, mas propõe alternativas para inibição das situações de risco

Comentários e Considerações sobre a Pesquisa:
Sem comentários

Considerações sobre os Termos de apresentação obrigatória:
Todos os termos foram devidamente apresentados. O cronograma prevê a realização da captura em 20 de abril. Contudo, o proponente inclui documento de atestado ético quanto ao início da pesquisa após aprovação pelo CEP

Recomendações:
Sem recomendações

Endereço: Rua: Dr. Carlos Chagas, s/ n
Bairro: São José **CEP:** 58.107-670
UF: PB **Município:** CAMPINA GRANDE
Telefone: (83)2101-5545 **Fax:** (83)2101-5523 **E-mail:** cep@huac.ufcg.edu.br

UFCG - HOSPITAL UNIVERSITÁRIO ALCIDES CARNEIRO DA UNIVERSIDADE

Conclusões ou Pendências e Lista de Inadequações:

Sem considerações finais

Considerações Finais a critério do CEP:

Parecer aprovado em reunião realizada em 23 de abril de 2018.

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1076489.pdf	12/04/2018 10:39:29		Aceito
Declaração de Pesquisadores	Termo_de_divulgacao_dos_resultados.pdf	12/04/2018 10:38:07	SERGIO CAVALCANTI DE PAIVA	Aceito
Declaração de Pesquisadores	Declaracao_informando_iniciara_a_coleta_apenas_apos_aprovacao_do_CEP.pdf	12/04/2018 10:36:19	SERGIO CAVALCANTI DE PAIVA	Aceito
Projeto Detalhado / Brochura Investigador	PROJETO.pdf	06/04/2018 10:15:14	SERGIO CAVALCANTI DE PAIVA	Aceito
Folha de Rosto	folhaDeRostoAssinada.pdf	06/04/2018 10:14:23	SERGIO CAVALCANTI DE PAIVA	Aceito
Declaração de Pesquisadores	TermoCompromissoPesquisadores.pdf	06/04/2018 10:07:54	SERGIO CAVALCANTI DE PAIVA	Aceito
Declaração de Instituição e Infraestrutura	TERMO_DE_ANUENCIA_INSTITUCIONAL.pdf	06/04/2018 09:48:57	SERGIO CAVALCANTI DE PAIVA	Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TERMO_DE_CONSENTIMENTO_LIVRE_E_ESCLARECIDO.pdf	27/02/2018 01:05:46	SERGIO CAVALCANTI DE PAIVA	Aceito

Situação do Parecer:
Aprovado

Necessita Apreciação da CONEP:
Não

Endereço: Rua: Dr. Carlos Chagas, s/ n
Bairro: São José **CEP:** 58.107-670
UF: PB **Município:** CAMPINA GRANDE
Telefone: (83)2101-5545 **Fax:** (83)2101-5523 **E-mail:** cep@huac.ufcg.edu.br

UFCG - HOSPITAL UNIVERSITÁRIO ALCIDES CARNEIRO DA UNIVERSIDADE

Continuação do Parecer: 2.618.913

CAMPINA GRANDE, 24 de Abril de 2018

Assinado por:
Januse Nogueira de Carvalho
(Coordenador)

Endereço: Rua: Dr. Carlos Chagas, s/ n
Bairro: São José **CEP:** 58.107-670
UF: PB **Município:** CAMPINA GRANDE
Telefone: (83)2101-5545 **Fax:** (83)2101-5523 **E-mail:** cep@huac.ufcg.edu.br

COMITÊ DE ÉTICA EM PESQUISA COM SERES HUMANOS - CEP
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE - UFCG
HOSPITAL UNIVERSITÁRIO ALCIDES CARNEIRO - HUAC

DECLARAÇÃO DE APROVAÇÃO DE PROJETO

Declaro para fins de comprovação que foi analisado e aprovado neste Comitê de Ética em Pesquisa – CEP o projeto de número CAAE: 87510318.5.0000.5182, Número do Parecer: 2.618.913 intitulado: **Sumarização de Vídeos à partir de Dados Fisiológicos e Expressões Faciais.**

Estando o (a) pesquisador (a) ciente de cumprir integralmente os itens da Resolução nº. 466/ 2012 do Conselho Nacional de Saúde – CNS, que dispõe sobre Ética em Pesquisa envolvendo seres humanos, responsabilizando-se pelo andamento, realização e conclusão deste projeto, bem como comprometendo-se a enviar por meio da Plataforma Brasil no prazo de 30 dias relatório do presente projeto quando da sua conclusão, ou a qualquer momento, se o estudo for interrompido.

Andréia Oliveira Barros Sousa
Coordenadora *pro tempore* CEP/ HUAC

Campina Grande - PB, 13 de Agosto de 2018.

Rua.: Dr. Carlos Chagas, s/ n, São José, Campina Grande – PB.
Telefone.: (83) 2101 – 5545. E-mail.: cep@huac.ufcg.edu.br

Anexo II

Artigos Apresentados

Neste anexo, se encontram dois artigos científicos derivados da tese, os quais foram aceitos em evento internacional (*IEEE 15th International Conference on Intelligent Computer Communication and Processing* (ICCP 2019)[1], Qualis B1).

1. *Investigation of Automatic Video Summarization using Viewer's Physiological, Facial and Attentional Features*: (PAIVA; GOMES, 2019a); e
2. *Link between Facial Expression and Emotional State Induced by Exposure to Multimedia Content*: (PAIVA; GOMES, 2019b).

Cópias destes artigos se encontram nas próximas páginas.

[1] http://www.iccp.ro/iccp2019/

Investigation of Automatic Video Summarization using Viewer's Physiological, Facial and Attentional Features

Sérgio Cavalcanti de Paiva*† and Herman Martins Gomes*
*Unidade Acadêmica de Sistemas e Computação, Universidade Federal de Campina Grande (UFCG)
Av. Aprígio Veloso 882, 58429-900 Campina Grande, PB, Brasil
Emails: sergiocp@copin.ufcg.edu.br, hmg@computacao.ufcg.edu.br
†Unidade Acadêmica de Serra Talhada,Universidade Federal Rural de Pernambuco (UFRPE)
Av. Gregório Ferraz Nogueira, s/n, 56909-535 Serra Talhada, PE, Brasil
Email: paivasc@gmail.com

***Abstract*—Video summarization aims at the selection of a concise and representative set of keyframes or video segments that allows the identification of the video content. In either cases, traditional summarization techniques usually work by segmenting the video into shots, representing video frames as feature vectors of color, texture, audio, among other features, clustering frames with similar features and selecting most representative keyframes or segments, sometimes guided by a video-to-summary ratio target. The resulting summaries are typically subject independent and do not take into account specific viewer's behavior. Instead of using intrinsic features extracted from the video for summarization, in this article we study whether personalized (subject dependent) video summaries can be obtained from physiological, facial, and attentional data captured from the viewers. More specifically, we study the relationship between personalized video summaries reported by viewers and their data captured during the display of different video genres. A dataset of fifteen videos was used in the experiments. During the exhibition of the videos, the viewer's physiological, facial, and attentional data were recorded, analyzed and synchronized. Several machine learning models were trained to test our hypothesis. We obtained k-fold cross validation accuracies that were above the chance for the best learned models. As a result of this study, we conclude that it is possible to train a learning machine that can produce customized summaries that are closer to user preferences compared to randomly produced summaries.**

I. INTRODUCTION

Among the strategies for large-scale video processing, summarization is considered a milestone for filtering video contents that is most relevant to consumers. Video summaries might be integrated with functions such as search engines and interactive video browsing.

Video summarization techniques may be classified into three categories related to the features used to generate the summary: internal, external and hybrid. Techniques that use information derived directly from the contents of the video stream (image, sound and text) are called internal summarization techniques. These are distinct from external techniques that take into account information collected externally from the video stream (e.g. information entered by the user or acquired from the environment). And hybrid summarisation use both internal and external sources [1].

The idea of generating video summaries based on the perception of the viewers emerges as an important research topic. This can be accomplished by means of high level concepts obtained from an affective analysis of the video contents. Such approach is closely connected with the participant's attention and emotions and poses a promising direction in the quest to reduce the semantic gap between low-level features extracted from the video and the high-level concepts perceived by the viewers with the aim of creating customized summaries [2].

In this paper we analyzed video summarization strategies supported by machine learning techniques. The core data for the analysis was physiological, facial and attentional features acquired from the viewers during sessions of video exhibition. After being presented with some video contents, viewers were asked to select keyframes that should be part of a video summary.

The present study was guided by the following research question: **Is it possible to obtain automatic video summarizers based on monitored physiological, facial, and attentional data of viewers?** From this main question the following secondary question was also formulated: *How much better are the automatic video summaries compared to random summaries?*

II. RELATED WORK

Humans expose their feelings unconsciously by various means, such as facial expressions, behavioral responses and eye movements, in response to stimuli of various natures, such as visual and auditory. In this sense, viewers watching videos also exhibit such rich information related to experienced sensations during exposition to the videos.

Research in the field of monitoring users' spontaneous physiological and behavioral responses as a means to understand emotional processing of videos is recent. Emotional

978-1-7281-4914-1/19/$31.00

labeling of videos could be automatically accomplished by a recognition process that maps physiological or behavioral cues into emotional categories during the display of videos [3].

One of the pioneers to investigate whether physiological responses from participants can be used to provide summaries of affective content from videos were Money and Agius [1], who analyzed electro-dermal responses (EDR), respiration amplitude (RA), respiration rate (RR), blood volume pulse (BVP) and heart rate (HR) as a way to detect the most relevant sub-segments of videos to a given viewer. A variety of video genres was considered in their research, such as horror, comedy, drama, sci-fi and action.

These physiological measures were considered by Money and Agius [4] in the proposition of the ELVIS technique (Entertainment-Led VIdeo Summaries), which was used to process and analyze physiological response data and to identify the sub-segments that were more entertaining, according to the user's physiological responses. Further experiments using the ELVIS technique with a larger set of users and additional genres was presented in a later work [5].

An approach to the detection of personal highlights and video summarization based on the analysis of the viewers' facial activity was proposed in the articles of Joho et al. [6] and Joho et al. [7].

In the first article, Joho et al. [6] proposed two models: pronounced level, and expression's change rate. The first model was motivated by the observation that certain facial expressions are more pronounced than others. The second model was related to the expression's change rate from one category to another. While the frequency of changes was treated as analogous to the change that occurs in the affective state of the viewer, this model counts the number of frames where the same category remains dominant in each frame. These models were then combined.

In the second article, Joho et al. [7] have developed a real-time facial expression recognition system that is composed of a face tracking algorithm that generates a vector of motion features of certain face regions that are feed into a Bayesian network classifier. The main idea was to detect personal highlights of multimedia contents based on these features.

The research of Peng et al. [8] proposed to perform video editing by viewing. For this, the authors analyzed the blinks, saccades, head motions and facial expressions of the viewer while watching a video. From these analyzes, models of attention and emotion were generated, which, in turn, were employed to estimate an Interest Meter (IM) relative to the viewers, based on their behavior during exposition to video contents.

III. MATERIALS AND METHODS

A. Overview

A video dataset of fifteen videos was obtained from the YouTube platform (see section III-D). The videos were segmented into shots and each video shot was represented by a keyframe. We also obtained physiological, visual attention, and facial expression data of the viewers when watching the videos. The physical structure of the data acquisition system is shown in Figure 1. A neutral video was displayed before each video exhibition. The average duration of the experiments was approximately one and a half hours. Thirty-three healthy participants (twelve women and twenty-one men, with ages ranging from 18 to 48 years) participated in the experiment.

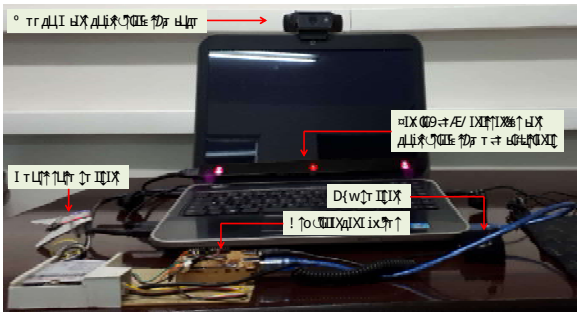


Fig. 1. Physical structure of the data acquisition system

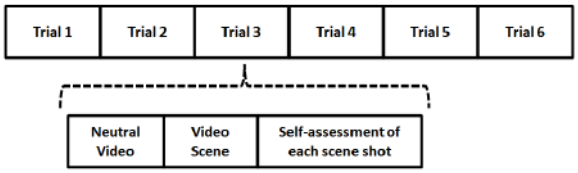


Fig. 2. Scheme of the experimental protocol used, with details of the trial

After finishing the experiment, two types of affective information about each video were available: (i) Information obtained from the participants: physiological, attention and Action Units (AUs) data extracted from the participant's face, and (ii) Identification (by the viewers) of the video shots that should belong to the summary, used as 'ground truth' for the viewer's preferences in each video shot.

B. Experiments

In this work, each data acquisition experiment was split into six trials and planned to last one and a half hours, on average. Each trial consisted of the exhibition of a neutral video, followed by the exhibition of a video from the dataset. After that, the viewer's preferred frames (based on video shots) was acquired (Figure 2). This protocol was designed to reduce risk of fatigue by the participants and was approved by the Ethics Committee of the University where this research was conducted. Moreover, anonymity was assured to all participants.

Before the trials, each participant received a brief explanation of the research, its expected duration and the types of sensors used for data acquisition. Participants were asked about the environmental conditions of the acquisition so that eventual adjustments could be made beforehand.

C. Trials

The trials were organized in three parts:

TABLE I
YOUTUBE VIDEOS USED IN THE EXPERIMENTS

YouTube ID[a]	Genre	Start	Duration
ep6ZtCmsK7I	comedy	00:00:00	00:02:12
fw8gKiaXi7c	comedy	00:00:00	00:01:42
l3SKZK_m9n4	comedy	00:00:00	00:01:24
xN_9GSBIDf8	comedy	00:00:00	00:02:02
ii7FL7t7I98	comedy	00:00:34	00:02:29
gozRrRCtj6E	horror/thriller	00:00:00	00:01:58
0mkMFHkhcTo	horror/thriller	00:00:24	00:02:25
toAOUXtlXXc	horror/thriller	00:00:00	00:02:13
0wM1nNx7t4A	horror/thriller	00:00:06	00:01:07
E6JnsVZ2MRE	horror/thriller	00:00:00	00:01:05
brUjCnSv0no	tragedy/drama	00:00:06	00:02:28
xXLoEaCfDV0	tragedy/drama	00:00:00	00:02:37
c3B18awVWcY	tragedy/drama	00:00:03	00:01:15
yztusel4vFw	tragedy/drama	00:01:20	00:02:19
teJzjgXIP3k	tragedy/drama	01:20:24	00:02:36

[a]The complete address is formed as https://www.youtube.com/watch?v=ID

1) Display of neutral videos to assist in eliminating the influence of previous videos;
2) Participant's data acquisition during video displays:
   - physiological: skin condutance (GSR) and heart rate (HR) via sensors connected to an Arduino computer;
   - eye fixations: (x,y) screen coordinates where participants fixated, using Tobii EyeX Controller[2], and
   - video of the participant's face, with the purpose of facial expression analysis using OpenFace tool.
3) Selection, by the participant, of the keyframes that should belong to a personalized summary, when considering 15 to 30% of the total number of video frames.

D. Data

1) Video dataset: The videos selected from YouTube came from the various genres: comedy, tragedy/drama and horror/thriller.

The complete video dataset was composed of five videos of each genre. Two random videos from each genre were chosen for display to a particular participant, who viewed six videos in total. The length of the videos ranged from one minute and five seconds to two minutes and thirty-seven seconds. Table I lists the selected videos with their locations, genres, start time and durations.

E. Data synchronization

The features obtained from the sensors are stored in textual and multimedia files. After the data acquisition phase, data synchronization takes place. This phase starts with creating a record of fixations. The time intervals of the fixations are obtained from the intervals between a "begin" and an "end" of the record containing eye fixation. Fixation features were obtained for the durations of 0.5, 1.0, 1.5 and 2.0 seconds.

Fixation and physiological data are synchronized using the start time of each video frame as reference. Videos of the participant's faces are processed through the OpenFace tool [9] in order to have the *Action Units* features extracted. As a result, a textual record containing features for each frame of the participants video. These two records are then fused and, afterward, a global synchronization record is created, which refers to the displayed video shots intervals. Additionally, the duration and number of the fixations are added the this global synchronization record. More details about the action units are presented in the next paragraphs.

1) Action Units (AUs): Videos containing the face of the participant were processed by means of OpenFace 2.0 toolkit [10]. This toolkit supports facial behavior analysis experiments and extracts many types of features relating to head pose, eye-gaze, AUs, among others. Only features relating to the intensity of the AUs, recorded frame-by-frame, were taken into consideration. Face detection and alignment are performed via structural SVM followed by Convolutional Experts Constrained Local Model (CE-CLM). Corrections in the face plane orientation are also performed. Histograms of oriented gradients (HOGs) [11] produce appearance features. The OpenFace toolkit provides AU intensity estimation through Support Vector Regression (SVR), and AU detection through Support Vector Machine (SVM), in both cases using linear kernels. Since the AUs occurrence is naturally unbalanced, a sub-sampling of negative AU samples from the training data was performed, with the aim to produce an equal number of positive and negative samples [12], [10].

The subset of AUs recognized by OpenFace are: 1, 2, 4, 5, 6, 7, 9, 10, 12,14, 15, 17, 20, 23, 25, 26, 28 and 45, according to the Facial Action Coding System (FACS). The FACS is a system based on the anatomical knowledge of facial muscles and their configurations of movements for measuring facial behavior, initially proposed by Ekman and Friesen in their work entitled *"Universal and cultural differences in facial expression of emotion"* [13]. FACS relates to the small face movements that result from emotions and other physiological states.

F. Machine Learning Models

In order to study the relationship between the viewers's data and the self-reported summaries, six different machine learning models and three re-sampling strategies were employed: logistic regression (LogitBoost), linear regression (pls and simpls), naive_bayes, decision trees (Random Forest) and Support Vector Machine.

Since the number of frames belonging to the abstract is much smaller than those that should not belong, this is a set of data that negatively affects the performance of the learning machines, and thus must be treated with resampling strategies. The following resampling strategies were used: *Under*, *SMOTE* and *ROSE*. According to [14], the Under strategy performs the subsample of the majority class by creating a random subset of the data of this class to match the population of the minority class and the population of this class is maintained, and the SMOTE (Synthetic Minority Oversampling TEchinque) strategy creates a data set in which

[2]https://www.tobii.com/

the minority class is augmented, creating "synthetic" data, interpolating values between an instance and nearest randomly selected neighbors, and reducing the majority class. The ROSE (Random Over-Sampling Examples) strategy, according to [15], implements a strategy similar to SMOTE, but the "synthetic" data is evenly distributed in the neighborhood of the minority class. For experimental evaluations, the selected keyframes (that should belong to a summary) are the outputs.

These models are implemented in the **caret** package, as part of the open source statistical environment R[3] [16]. Caret has several functions focused on simplifying the process of constructing and evaluating complex classification and regression models, as well as the selection of features and other techniques [17].

For training, we focused on optimizing classification accuracy. We estimated the performance of a given model using repeated n-fold cross validation, with 10 folds and 3 replicates. Parameter tuning was handled automatically by the **caret** package.

IV. RESULTS AND DISCUSSIONS

For each participant, the records of all viewed videos were first grouped by obtaining a vector of features $\Phi = \{F_1, F_2, \ldots, F_i\}$, where F_i is the feature set of the video i and represented as:

$$F_i = \{f(i,1), f(i,2), \ldots, f(i,\lambda)\} \text{ and}$$
$$f(i,j) = \{f^1_{(i,j)}, f^2_{(i,j)}, \ldots, f^\eta_{(i,j)}\}$$

where λ is the number of keyframes belonging to the video i, $f(i,j)$ is the set of features belonging to the key frame j of the video i, $f^l_{(i,j)}$ is the value of the feature l for the key frame j of the video i and η is the number of features used in this research, which was twenty-five (25), which are the physiological features (two features), Action Units (seventeen features) and fixations (six features).

For each feature of the vector Φ was applied to Zero-mean normalization, which follows the equation IV.

$$\check{f}^l_{(i,j)} = \frac{f^l_{(i,j)} - \bar{f}^l}{\sigma}$$

where $\check{f}^l_{(i,j)}$ is the normalized feature value l, $\bar{f}^l$ is the feature mean l and σ is its standard deviation of feature l. Thus, we also create the value for the normalized feature vector ($\check{\Phi}$), as shown in the definitions below:

$$\check{f}(i,j) = \{\check{f}^1_{(i,j)}, \check{f}^2_{(i,j)}, \check{f}^3_{(i,j)}, ..., \check{f}^\eta_{(i,j)}\}$$
$$\check{F}_i = \{\check{f}(i,1), \check{f}(i,2), \ldots, \check{f}(i,\lambda)\}$$
$$\check{\Phi} = \{\check{F}_1, \check{F}_{(2}, \ldots, \ldots, \check{F}_i\}$$

where $\check{f}(i,j)$ is the set of normalized features belonging to the key frame j of the video i and $\check{F}_i$ is the normalized feature set of the video i.

[3]https://www.r-project.org/

From the video groupings, 20 training and testing sets were generated. The usual proportion of 60-40% for training-testing was adopted. All training sets have been used for training combinations of the six learning machines with the three re-sampling strategies. Two training variations were investigated:

- utilizing all features obtained from the participant (ALL) and
- utilizing only the physiological features and facial expressions (SF).

The method used to evaluate each summary was the one proposed by Avila et al. [18], known as Comparison of User Summaries (CUS). In this method, the quality of the self-generated summary is determined by two metrics, named CUS_A precision rate and CUS_E error rate, which are defined as:

$$CUS_A = \frac{n_{mAS}}{n_{US}}$$

$$CUS_E = \frac{n_{\bar{m}AS}}{n_{US}}$$

where n_{mAS} is the number of corresponding keyframes in the automatic summary (AS), $n_{\bar{m}AS}$ is the number of non-matching keyframes from the AS, and n_{US} is the number of keyframes of the user summary (US). As the quantity of frames for the summary is determined a priori, the values of CUS_A and CUS_E become complementary, then the following results were analyzed considering the values of CUS_A.

In order to verify the efficiency of the learned models we compared them with chance (RANDOM). Self-reported frames of each participant were used as ground truth for the summaries. Consequently, the following hypothesis is proposed: *CUS_A of the random selection, on average, is greater than or equal to the CUS_A of the learning machine selection (with re-sampling strategy only in the training set)* . To test the hypothesis, the Student's t-test comparisons were applied when both selections followed normal distribution. The Wilcoxon's median test was applied, otherwise. In both cases, the tests were performed for paired samples. When p-value is less than the significance level of 0.05, we reject the null hypothesis and accept the alternative hypothesis (*CUS_A of random selection is less than the CUS_A of machine selection*).

The training options ALL and SF, when choosing a global combination of learning machine and re-sampling strategy (i.e. the same learning machine and re-sampling strategy is adopted for all participants), obtained 21 and 24 participants in which this combination obtained results superior to the random selection, respectively. The combinations that presented the best result when compared to the random selection obtained 24 and 26 participants in which this combination obtained results superior to the random selection, respectively. In the Tables II and III, the results of the best training option (BTO) per participant of the experiments and the best global combination (BGC), when it exists for the participant, are presented.

As it can be seen in the "p.value" column of the Tables II and III, these machines performed significantly better than

TABLE II

BEST SETS OF LEARNING MACHINE AND RE-SAMPLING STRATEGY BY PARTICIPANT WHOSE DATA FOLLOW NORMAL DISTRIBUTION. TRAINING OPTION BTO AND/OR BGC. COLUMNS CONTAIN PARTICIPANT ID, LEARNING MACHINE, RE-SAMPLING STRATEGY AND CORRESPONDING MEANS AND CONFIDENCE INTERVALS FOR THE CUS_A VALUES OF THE LEARNING MACHINES AND RANDOM SELECTION

					Random		Learning Machine	
ID	Type*	Machine	re-sampling	p.value	$\overline{CUS_A}$	Confid. Interval	$\overline{CUS_A}$	Confid. Interval
03	BTO	rf	Under	0.00	0.26	[0.21; 0.30]	0.43	[0.38; 0.47]
04	BTO	svmRadial	SMOTE	0.01	0.24	[0.20; 0.28]	0.31	[0.28; 0.34]
07	BTO and BGC	LogitBoost	Under	0.00	0.22	[0.19; 0.26]	0.32	[0.28; 0.36]
12	BGC	LogitBoost	Under	0.00	0.26	[0.23; 0.29]	0.39	[0.35; 0.44]
	BTO	rf	SMOTE	0.00			0.40	[0.37; 0.43]
15	BTO and BGC	LogitBoost	Under	0.03	0.22	[0.18; 0.26]	0.29	[0.22; 0.35]
17	BGC	LogitBoost	Under	0.00	0.28	[0.24; 0.31]	0.42	[0.38; 0.47]
	BTO	pls	Under	0.00			0.40	[0.38; 0.42]
20	BTO and BGC	LogitBoost	Under	0.00	0.19	[0.16; 0.22]	0.26	[0.23; 0.30]
21	BGC	LogitBoost	Under	0.00	0.23	[0.2; 0.26]	0.38	[0.34; 0.41]
	BTO	svmRadial	SMOTE	0.00			0.38	[0.35; 0.41]
22	BGC	LogitBoost	Under	0.00	0.33	[0.29; 0.37]	0.43	[0.39; 0.47]
	BTO	pls	Under	0.00			0.46	[0.42; 0.49]
23	BTO and BGC	LogitBoost	Under	0.00	0.31	[0.28; 0.34]	0.38	[0.34; 0.41]
25	BTO	rf	SMOTE	0.00	0.15	[0.11; 0.19]	0.24	[0.21; 0.27]
26	BGC	LogitBoost	Under	0.00	0.18	[0.13; 0.22]	0.34	[0.29; 0.39]
	BTO	svmRadial	SMOTE	0.00			0.32	[0.28; 0.37]
27	BGC	LogitBoost	Under	0.00	0.19	[0.16; 0.22]	0.30	[0.25; 0.35]
	BTO	naive_bayes	SMOTE	0.00			0.32	[0.29; 0.34]
28	BGC	LogitBoost	Under	0.01	0.21	[0.16; 0.25]	0.30	[0.25; 0.35]
	BTO	svmRadial	SMOTE	0.00			0.38	[0.34; 0.41]
29	BTO and BGC	LogitBoost	Under	0.00	0.23	[0.19; 0.26]	0.31	[0.27; 0.35]
30	BGC	LogitBoost	Under	0.00	0.21	[0.17; 0.25]	0.30	[0.25; 0.34]
	BTO	svmRadial	SMOTE	0.00			0.33	[0.31; 0.36]

* This column indicates the Best Training Option found (BTO), and the result found for the Best Global Combination (BGC) that was superior to that of the random selection.

TABLE III

BEST SETS OF LEARNING MACHINE AND SAMPLING STRATEGY BY PARTICIPANT WHOSE DATA DOES NOT FOLLOW NORMAL DISTRIBUTION. TRAINING OPTION IS BTO OR BGC. COLUMNS INDICATE PARTICIPANT ID, LEARNING MACHINE, RE-SAMPLING STRATEGY, AND CORRESPONDING MEDIANS, 1^{st} QUARTILE E 3^{rd} QUARTILE FOR CUS_A VALUES OF LEARNING MACHINES AND THE RANDOM SELECTION

					Random		Learning Machine	
ID	Type	Machine	re-sampling	p.value	CUS_A	$[1^{st}q; 3^{rt}q.]$	CUS_A	$[1^{o}q; 3^{o}q.]$
01	BGC	LogitBoost	Under	0.01	0.32	[0.28; 0.33]	0.38	[0.28; 0.48]
	BTO	svmRadial	SMOTE	0.00			0.40	[0.35; 0.41]
03	BTO and BGC	LogitBoost	Under	0.00	0.28	[0.20; 0.30]	0.40	[0.34; 0.50]
05	BGC	LogitBoost	Under	0.00	0.29	[0.24; 0.30]	0.40	[0.33; 0.48]
	BTO	simpls	Under	0.00			0.38	[0.33; 0.39]
06	BTO and BGC	LogitBoost	Under	0.00	0.25	[0.23; 0.30]	0.33	[0.30; 0.37]
08	BTO and BGC	LogitBoost	Under	0.02	0.17	[0.13; 0.19]	0.19	[0.19; 0.25]
10	BGC	LogitBoost	Under	0.03	0.24	[0.24; 0.29]	0.31	[0.24; 0.35]
	BTO	svmRadial	SMOTE	0.01			0.32	[0.29; 0.35]
11	BGC	LogitBoost	Under	0.00	0.14	[0.12; 0.16]	0.26	[0.20; 0.28]
	BTO	LogitBoost	ROSE	0.00			0.24	[0.20; 0.29]
13	BTO and BGC	LogitBoost	Under	0.00	0.16	[0.14; 0.23]	0.27	[0.18; 0.38]
19	BGC	LogitBoost	Under	0.00	0.30	[0.23; 0.31]	0.43	[0.40; 0.49]
	BTO	rf	SMOTE	0.00			0.41	[0.40; 0.46]
24	BGC	LogitBoost	Under	0.00	0.19	[0.15; 0.20]	0.27	[0.22; 0.38]
	BTO	svmRadial	SMOTE	0.00			0.31	[0.26; 0.31]
32	BTO and BGC	LogitBoost	Under	0.01	0.25	[0.25; 0.29]	0.31	[0.27; 0.41]

* This column indicates the Best Training Option found (BTO), and the result found for the Best Global Combination (BGC) that was above that of the random selection.

TABLE IV

PARTICIPANTS THAT OBTAINED CUS_A BETTER THAN THAT FROM RANDOM SELECTION. BTO=BEST TRAINING OPTION, BGC=BEST GLOBAL COMBINATION.

ALL	BGC	03, 05, 07, 10, 11, 12, 13, 17, 18, 19, 21, 22, 23, 24, 26, 27, 28, 29, 30, 32
	BTO	03, 04, 05, 07, 10, 11, 12, 13, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32
SF	BGC	01, 03, 05, 06, 07, 08, 10, 11, 12, 13, 15, 17, 19, 20, 21, 22, 23, 24, 26, 27, 28, 29, 30, 32
	BTO	01, 03, 04, 05, 06, 07, 08, 10, 11, 12, 13, 15, 17, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32

RANDOM with significance level of 5% for these participants. Thus, we can say that in most cases "it is possible to obtain automatic summarizers of videos from the physiological, facial and monitored characteristics of viewers".

The results obtained in the Tables II and III can be summarized as:

- For most participants, it is possible to train a learning machine (with a re-sampling strategy only in the training set) that can predict (with higher correct decision and lower error rates compared to random decisions) whether

a given video keyframe (whose physiological data of the participant were never seen by the learning machine) should or should not be part of a video summary of a predetermined size,

- If the participant analyzes in the experiment were done by choosing a global combination of learning machine and re-sampling strategy, the participants with results that were better than the random selection would belong to a smaller group than when using specific (participant-dependent) combinations of learning machines and re-sampling strategy. This leads us to believe that the choice of the pair "learning machine" and "re-sampling strategy" should also be individualized per participant.
- The difference between the training option ALL with the SF option are the participants 16 and 18, and the difference between the training option SF with the ALL option are the participants 1, 6, 8 and 15, as it can be observed in Table IV. This indicated that the information resulting from the eye fixation for two participants was relevant, however, for most participants, it was irrelevant, or for the case of these four specific participants, eye fixation was worsening the results.

V. Conclusions

As a response to the research question, this study corroborates the idea that automatic video summarizers can be obtained from physiological, facial and attentional features captured from viewers. When analyzing the accuracy rates CUS_A obtained for the participants, it was verified that, in their majority, these accuracies have exceeded the ones obtained by a random selection of keyframes for the summaries. Since most related work in automatic video summarization using features extracted from the viewers does not perform objective evaluations (e.g. [5], [8] and [19]), it is difficult to make a direct comparison of our approach with those works. However, the results presented in Table II corroborate those obtained by Money and Agius [4] in their study that used physiological features to identify the most entertaining video sub-segments (self-reported by the viewers).

In this work, we contributed with an approach for performing a personalized selection of the most representative shots of a video, represented by its keyframes. An objective evaluation was also carried out, in which automatic personalized summaries were compared to previous self-reported summaries from the corresponding participants of the experiment. This differs from what is found in the literature of the area, where participants are asked to give a score to the automatic/personalized summaries after they are created.

Besides the complexity of investigating such a subjective theme that is the human behavior, other difficulties faced when performing this study were related to the recruitment of a sufficient number of volunteers to participate in the experiments and the selection of video content that was suitable to test our hypothesis.

As for future work, we propose to carry out experiments by expanding the experimental protocol, with an increase in the number of videos, and usage of video information, such as the video genre, as input to the machine learning models in order to assess eventual improvements in model accuracy. Moreover, analyses based on viewers' age and gender will allow testing more specific hypothesis.

References

[1] A. G. Money and H. Agius, "Analysing user physiological responses for affective video summarisation," *Displays*, vol. 30, no. 2, pp. 59 – 70, 2009.

[2] ——, "Video summarisation: A conceptual framework and survey of the state of the art," *Journal of Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121 – 143, 2008.

[3] S. Wang, Y. Zhu, G. Wu, and Q. Ji, "Hybrid video emotional tagging using users' EEG and video content," *Multimedia Tools and Applications*, vol. 72, no. 2, pp. 1257–1283, apr 2013.

[4] A. G. Money and H. Agius, "Elvis: Entertainment-led video summaries," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 6, no. 3, pp. 17:1–17:30, Aug. 2010.

[5] ——, "'mind the gap': Evaluating user physiological response for multi-genre video summarisation," in *Proc.*, ser. BCS-HCI '13, International BCS Human Computer Interaction Conference. Swinton, UK, UK: British Computer Society, 2013, pp. 37:1–37:6.

[6] H. Joho, J. M. Jose, R. Valenti, and N. Sebe, "Exploiting facial expressions for affective video summarisation," in *Proc.*, ser. CIVR '09, ACM International Conference on Image and Video Retrieval. New York, NY, USA: ACM, 2009, pp. 31:1–31:8.

[7] H. Joho, J. Staiano, N. Sebe, and J. M. Jose, "Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents." *Multimedia Tools Appl.*, vol. 51, no. 2, pp. 505–523, 2011.

[8] W. Peng, W. Chu, C. Chang, C. Chou, W. Huang, W. Chang, and Y. Hung, "Editing by viewing: Automatic home video summarization by viewing behavior analysis," *IEEE Transactions on Multimedia*, vol. 13, no. 3, pp. 539–550, 2011.

[9] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, mar 2016.

[10] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, may 2018.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, sep 2010.

[12] T. Baltrusaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, may 2015.

[13] A. Hildebrandt, S. Olderbak, and O. Wilhelm, "Facial emotion expression, individual differences in," in *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, 2015, pp. 667–675.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, jun 2002.

[15] N. Lunardon, G. Menardi, and N. Torelli, "ROSE: a package for binary imbalanced learning," *The R Journal*, vol. 6, no. 1, p. 79, 2014.

[16] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: http://CRAN.R-project.org/doc/Rnews/

[17] M. Kuhn, "Building predictive models in r using the caret package," *Journal of Statistical Software*, vol. 28, no. 5, 2008.

[18] S. E. F. de Avila, A. P. B. a. Lopes, A. da Luz, Jr., and A. de Albuquerque Araújo, "Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method," *Pattern Recognition Letters*, vol. 32, no. 1, pp. 56–68, jan 2011.

[19] I. Mehmood, M. Sajjad, S. W. Baik, and S. Rho, "Audio-visual and EEG-based attention modeling for extraction of affective video content," in *2015 International Conference on Platform Technology and Service*. IEEE, jan 2015.

Link between Facial Expressions and Emotional States Induced by Exposure to Multimedia Content

Sérgio Cavalcanti de Paiva*† and Herman Martins Gomes*
*Unidade Acadêmica de Sistemas e Computação, Universidade Federal de Campina Grande (UFCG)
Av. Aprígio Veloso 882, 58429-900 Campina Grande, PB, Brasil
Emails: sergiocp@copin.ufcg.edu.br, hmg@computacao.ufcg.edu.br
†Unidade Acadêmica de Serra Talhada,Universidade Federal Rural de Pernambuco (UFRPE)
Av. Gregório Ferraz Nogueira, s/n, 56909-535 Serra Talhada, PE, Brasil
Email: paivasc@gmail.com

***Abstract*—The explosive growth of digital videos has created new challenges for computer science. While many advances on video indexing, retrieval and summarization based on general, subject-independent, objective descriptors have been made in the past years, research on the use of individual subjective preferences and affective states is at the forefront of research and poses great challenges. In this article, we study the relationship between emotional states reported by viewers and their facial physiological changes observed during the display of different video genres. A dataset of twenty videos was created from YouTube video sharing platform. During the exhibition of the videos, the viewer's facial activities have been recorded and analyzed by means of Action Units (AUs). After that, emotional states self-reported by the viewers were assigned to video shots. Labels were divided into four categories, defined according to a discrete version of Russel's Circumplex emotion model. Different machine learning models were trained to test the relationship between the measured facial features and the self-reported emotional categories. We obtained k-fold cross validation accuracies that were above chance for the best learned models. As a result of this study, we concluded that AUs can indeed be used as an valuable tool to estimate emotional categories during exposure to audiovisual stimuli, and, therefore, should be used in further studies that take advantage of those categories to devise personalized multimedia retrieval and summarization approaches.**

I. INTRODUCTION

Digital videos are increasingly present in our lives. This creates the need for more effective technologies for the management, storage and transmission of digital video contents [1]. In this way, new multimedia indexing and retrieval methods become indispensable for managing and understanding video contents.

Research into the topic of video affective content analysis has attracted a lot of attention in the last few years, as viewers can feel a wide variety of emotions and anticipate those that a movie may provoke. Such information can be highly beneficial not only for improved accuracy in video indexing and compression, but also for the delivery of customized content based on the viewer‘s mood [2].

In this article, we investigate whether emotions felt by viewers during the exposition videos of different genres could be linked to some physiological cues. Facial expressions produced during video exposure were measure by means of Action Units [3]. Viewers provided an emotional score related to each video shot, represented as a keyframe.

More specifically, this research was guided by the following research question: **Can components of facial expression (Action Units - AUs) be used to estimate self-reported emotional categories during exposure to audio-visual stimuli?** The secondary issues were formulated from this main question: *Can a reduced subset of AUs produce better classification results than those produced when using all AUs? Are classification results worse when you restrict training and test data from coming from different videos?*

The remaining of this article is organized as follows. Next section discusses the area of affective video analysis. Section III presents the materials and methods adopted in the investigation. Results and discussions are presented on Section IV. Finally, the conclusions and some proposals of future work are provided in Section V.

II. AFFECTIVE VIDEO ANALYSIS

One of the goals of affective video analysis is to automatically recognize the emotions evoked in the viewers after exposition to their contents. There are three perspectives for analysis related to specific emotion detection: those induced by the film maker to the viewers (intended emotion); those felt by viewers in response to the video (induced emotion); and the emotions felt by the majority of the audience in response to the same content (expected emotion) [4]. The answers obtained from each of these perspectives do not always match. In this sense, we seek answers that will allow personalized affective indexing of video scenes based on spectator's induced emotion.

Before performing affective video analysis, it is necessary to define the approach used to describe the emotions. Among all existing approaches, the most used for affective analysis of multimedia content are the categorical and dimensional approaches [5]. In the discrete (or categorical) approach, emotion is labeled in several discrete categories, whereas in the dimensional approach, emotion is represented in a continuous space [6]. In this way, the approaches internally present some

978-1-7281-4914-1/19/$31.00

variations, as discussed next. Different categorical sets of emotions are proposed in the literature [6], but the most frequent is the basic set of categories of Ekman [3], where the emotion is labeled in six categories, named happiness, sadness, surprise, disgust, anger and fear. In the dimensional approach, most agree that describing a subjective response using two or three dimensions is sufficient for its representation (Some examples: [7], [8] and [9]). However, there is no consensus on the dimension labels [5]. One of the possible sets is that represented with valence, arousal and control (dominance). Valence is characterized by continuous responses ranging from pleasant to unpleasant, whereas arousal represents the intensity of emotion, ranging from excited to calm. The third dimension (dominance) is used to distinguish those emotions that present the two previous dimensions with similar values. This dimension vary continuously, from the absence of control to total control [10].

As discussed earlier, categorical and dimensional emotional descriptors have often been employed to describe the affective aspect of a video. However, both have their limitations, since emotions are complex and subjective. On the one hand, a limited number of discrete categories may not represent the subtlety and complexity of affective states, on the other hand, representing continuous and absolute scores may not be significant enough due to the lack of standardization for subjective emotions [5].

In this work the circumplex model of affection of Russell [11], see Figure 1, was used, so that the viewers could provide an emotional score to each of the video shots. Despite of being a dimensional model, words are presented in the figure to help indicate the location of some basic emotions (happiness, anger, fear, sadness etc) as well as other affective states that may or may not be considered as emotions (sleepy, serene) [12]. In this article, the emotional scores labeled by the viewers were represented in a discretized representation of four quadrants Q1, Q2, Q3 and Q4 (Figure 1).

III. MATERIALS AND METHODS

A. Overview

A video dataset of twenty videos was obtained from the YouTube platform, see section III-D. The videos were segmented into shots and each video shot was represented by a keyframe. We also obtained data about the facial expressions of viewers during video exposure and emotional labels self-assigned by the viewers, shot by shot, as a result of their experiences. A neutral video was displayed before each video exhibition. The average duration of the experiments was approximately one and a half hours. Thirty-three healthy participants (twelve women and twenty-one men, with ages ranging from 18 to 48 years) participated in the experiment. Thus, after finishing the experiment two types of affective information about each video were available: (i) Action Units (AUs) extracted from the participant's face, and (ii) self-assessed emotional states, used as 'ground truth' for the viewer's emotions in each video shot.

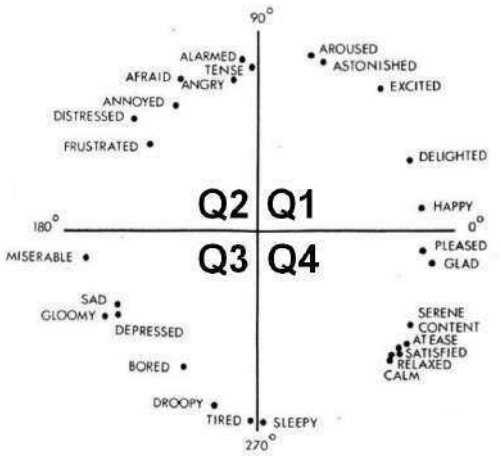


Fig. 1. Adaptation of Russell's circumplex model of affection [11]

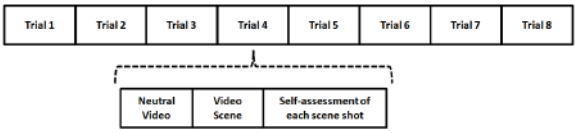


Fig. 2. Scheme of the experimental protocol used, with details of the trial

B. Experiments

Each data acquisition experiment was divided into eight trials and planned to have an average duration of one and a half hours. Each trial consisted of the exhibition of a neutral video, followed by the exhibition of a video from the dataset and the viewer's emotional self-assessment based on video shots (Figure 2). This experimental protocol was approved by the Ethics Committee of the University where the research was conducted and was designed to reduce the risk of fatigue by the participants.

Before the trials, participants were given instructions on how to proceed. In order to facilitate data collection, a brief explanation of the research was presented, with description of the object of study, its importance, the types of sensors used for data acquisition. Moreover, the participants received assurance of anonymity. Opinion about the ambient conditions (e.g. chair comfort, thermal, acoustic, illumination) was also inquired, so that necessary adjustments could be made beforehand.

C. Trials

The trials were organized in three parts: (1) normalization of emotions, with display of neutral videos to assist in eliminating previous emotional states; (2) Recording a video of the participants faces during the exhibition of videos; and (3) Labeling of the video shots.

TABLE I
YOUTUBE VIDEOS USED IN THE EXPERIMENTS

YouTube ID[a]	Genre	Start	Duration
ep6ZtCmsK7I	comedy	00:00:00	00:02:12
fw8gKiaXi7c	comedy	00:00:00	00:01:42
l3SKZK_m9n4	comedy	00:00:00	00:01:24
xN_9GSBIDf8	comedy	00:00:00	00:02:02
ii7FL7t7I98	comedy	00:00:34	00:02:29
gozRrRCtj6E	horror/thriller	00:00:00	00:01:58
0mkMFHkhcTo	horror/thriller	00:00:24	00:02:25
toAOUXtlXXc	horror/thriller	00:00:00	00:02:13
0wM1nNx7t4A	horror/thriller	00:00:06	00:01:07
E6JnsVZ2MRE	horror/thriller	00:00:00	00:01:05
brUjCnSv0no	tragedy/drama	00:00:06	00:02:28
xXLoEaCfDV0	tragedy/drama	00:00:00	00:02:37
c3B18awVWcY	tragedy/drama	00:00:03	00:01:15
yztusel4vFw	tragedy/drama	00:01:20	00:02:19
teJzjgXIP3k	tragedy/drama	01:20:24	00:02:36
CcLd0KtmUyI	neutral/calm	00:00:00	00:01:22
Bo-lvEodZfA	neutral/calm	00:00:00	00:02:38
wW8SciPlI0s	neutral/calm	00:00:00	00:01:49
aJT9F2oHrSg	neutral/calm	00:00:00	00:03:08
W3Ufvm7MqcM	neutral/calm	00:00:00	00:02:30

[a]The complete address is formed as https://www.youtube.com/watch?v=ID

D. Data

Video dataset: The videos selected from YouTube came from the following genres: comedy, tragedy/drama, horror/thriller and neutral/calm.

According to Xu et al. [13], the genres of action, horror, comedy and drama are among the most popular. Dominant emotions for horror and comedy movies are fear and happiness, respectivelly. For drama, however, it is difficult to name a dominant emotion, but this genre usually evokes many emotions. Action movies typically attract viewers attention and sustain emotions at high intensity most of the time. However, we decided not to use this last genre because it may lead to very distinct emotions such fear, anger and happiness. We added the neutral/relaxing genre to our dataset in order to both serve as a contrast and a control state for the other genres.

Each genre was represented by five videos in the dataset. Two random videos from each genre were chosen for displaying to a particular participant, who viewed eight videos in total. The length of the videos ranged from one minute and five seconds to three minutes and eight seconds. Table I contains a list of the selected videos with their locations, genres, start time and durations.

E. Feature Extraction

The OpenFace tool [14] processed the participants' face videos and obtained Action Units features, which are stored in a textual record for each frame of the participants' video. Subsequently, an average record of the features related to displayed video shots intervals.

Action Units (AUs): The videos containing the participant's faces were processed using the OpenFace 2.0 toolkit, described in Baltrusaitis et al. [14], and available for download on the github platform[2]. This toolkit is designed to support experiments on facial behavior analysis. From the rich record obtained by OpenFace 2.0, which includes, among others, head pose estimation, eye-gaze estimation and recognition of a subset of facial action units (AUs), only the characteristics related to the intensity of AUs recorded frame by frame in the video were considered in this research.

OpenFace performs face detection and alignment (frame-by-frame) using a structural SVM followed by Convolutional Experts Constrained Local Model (CE-CLM), which is an instance of a Constrained Local Model (CLM) followed by corrections related to the rotation of the face plane. This step generates 112 x 112 pixel images of the faces found. Appearance features are then extracted using Histograms of oriented gradients (HOGs), as in Felzenszwalb et al. [15]. The OpenFace toolkit provides AU intensity estimation, obtained via Support Vector Regression, and AU occurrence detection, obtained via Support Vector Machines. In both cases, linear kernels are employed. Aiming to balance the training data, since the occurrence of AUs is naturally unbalanced, a sub-sampling of negative AU samples from the training data was performed, leading to an equal number of positive and negative samples [16] and [14].

The subset of AUs recognized by the system are those with the id's 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26 and 45 on the Facial Action Coding System (FACS). The FACS is a system based on the anatomical knowledge of facial muscles and their configurations of movements for measuring facial behavior. Proposed by Ekman and Friesen in their work entitled *"Universal and cultural differences in facial expression of emotion"* [17], FACS relates to the small face movements that result from emotions and other physiological states.

F. Machine Learning Models

In order to study the relationship between the AUs and self-reported emotions, five different machine learning models were applied: K-Nearest Neighbor, Support Vector Machine, Random Forest, Neural Network and LogitBoost. For experimental evaluations, AUs (or a subset of those) are inputs to the models and the emotion label (quadrants Q1, Q2, Q3 or Q4) are the outputs.

These models are implemented in the **caret** package, as part of the open source statistical environment R[3] [18]. Caret (short for Classification And REgression Training) is a set of functions to simplify the modeling process for complex classification and regression problems. The package has several functions focused on simplifying the process of constructing and evaluating models, as well as the selection of features and other techniques [19]. The current version can be found in CRAN [4] and the project is hosted on the github platform[5].

[2]https://github.com/TadasBaltrusaitis/OpenFace
[3]https://www.r-project.org/
[4]http://CRAN.R-project.org/package=caret
[5]https://github.com/cran/caret

For training, we were interested in optimizing classification accuracy. We estimated the performance of a given model using repeated k-fold cross validation, with 10 folds and 3 replicates. Parameter tuning was handled automatically by the **caret** package.

G. Feature Selection

The relevance of the AU features related to the emotional scores obtained from the viewers, as presented in Section II was determined using the Boruta algorithm, available as a package of the R environment [18]. Boruta is a random forest based feature selection method, able to determine the relevance of features in an unbiased and stable way [20]. The essence of the algorithm is explained as follows. A random copy of the data is made, the copy is merged with the original and the classifier is constructed for this extended data [21]. The importance of the variable in the original data is evaluated when compared with that obtained by the random variables. Variables that have gained importance over the random variables are considered more relevant.

IV. RESULTS AND DISCUSSIONS

In this section, we show the results of the experiments carried out to address the research question and it's associated questions. The questions' responses aimed at comparing the results of machine learning with those of a random selection of emotional quadrants assigned to the frames provided by the each of the participants.

Action Units and the self-reported quadrants of the participants were used to train the five different machine learning. Machines with best accuracies were considered for analysis. Zero-mean normalization was applied to the input data as in equation 1.

$$x = \frac{x - \bar{x}}{\sigma} \tag{1}$$

where x is the normalized feature value, $\bar{x}$ is the feature mean and σ is its standard deviation.

For the first experimental evaluation, we selected the videos that presented the largest quantities of shots, labeled in each of the emotion quadrants, to be used in the classifier training. This way, we had four videos in the training set. The remaining videos were grouped and only those that contained labels from all four quadrants were used to form the test set. Data from participants with associated videos that did not meet the above requirements, were not used in the experiments. As a result, five participants (05, 08, 09, 12 and 26) had to excluded from the experimental evaluation. The average accuracies obtained for this experiment are shown in Table II. The highlighted values in the table are values that lie below the average accuracy of a random labeling of four classes (less than 0.25).

In order to compare the machine learning with the worst case baselines - keyframe selection by chance - thirty random keyframes selections were performed for the emotion labels, and the respective accuracies for each selection were calculated. The mean value for each participant was then calculated.

TABLE II
ACCURACIES FOR THE MACHINE LEARNING (KNN, SVM, RANDOM FOREST - RF, NNET E LOGITBOOST - LB) AND FOR THE RANDOM SELECTION BASELINE MODEL, ASSESSED BY PARTICIPANT. BOLDFACE VALUES INDICATE THE CASES WHERE A LEARNING MACHINE PERFORMED WORSE THAN THE RANDOM SELECTION.

Participant	KNN	SVM	RF	NNET	LB	Rand.
01	**0.191**	0.565	**0.243**	0.304	**0.198**	0.248
02	0.352	0.418	0.505	0.462	0.512	0.243
03	**0.264**	**0.250**	**0.181**	**0.236**	**0.148**	0.266
04	**0.195**	0.382	0.333	0.293	0.266	0.243
06	**0.123**	**0.216**	**0.136**	**0.123**	**0.245**	0.254
07	0.275	0.391	0.435	0.406	0.345	0.254
10	0.470	0.554	0.506	0.369	0.476	0.259
11	**0.240**	0.357	**0.170**	0.374	**0.222**	0.241
13	0.542	0.493	0.632	0.458	0.534	0.249
14	0.462	**0.246**	0.415	0.438	0.469	0.256
15	**0.236**	0.377	0.377	0.500	0.348	0.256
16	**0.209**	**0.235**	0.252	**0.243**	0.277	0.245
17	**0.248**	0.271	0.256	**0.241**	**0.244**	0.247
18	**0.070**	**0.174**	**0.243**	**0.122**	**0.221**	0.245
19	0.318	0.268	**0.255**	0.268	0.310	0.256
20	**0.235**	0.452	0.287	0.348	0.307	0.253
21	0.388	0.289	0.579	0.289	0.520	0.253
22	0.492	0.417	0.417	0.492	0.316	0.248
23	0.319	0.391	0.428	0.362	0.431	0.234
24	0.496	0.536	0.504	0.472	0.436	0.231
25	0.435	0.435	0.420	0.442	0.387	0.260
27	0.266	0.287	0.385	0.329	0.267	0.242
28	0.353	0.520	0.529	0.422	0.430	0.239
29	0.343	0.259	0.398	0.481	0.456	0.254
30	0.308	**0.243**	**0.225**	**0.219**	**0.178**	0.253
31	0.421	0.352	0.484	0.415	0.496	0.250
32	0.634	0.611	0.573	0.626	0.591	0.250
33	0.344	0.281	0.312	**0.250**	**0.242**	0.254

The accuracies obtained when using all available features were, for most cases, bigger than the average accuracies of the random selection. In relation to the participants, the lowest obtained mean accuracy was for 18, followed in ascending order by the accuracies for 06, 03, 30, 16 and 17, the latter with an average accuracy of just over 0.25.

In order to verify the question "facial expression components (AUs) can be used to classify self-reported emotional categories during exposure to audiovisual stimuli", we proposed the following null hypothesis: *random selection has a accuracy that is greater than or equal to the accuracy of the learning machine*. To test this hypothesis, when both selections follow normal distribution, we applied the Student's t-test comparison. Otherwise, the Wilcoxon's median test was applied. In both cases, these tests were performed for paired samples. In these tests, when p-value is less than the significance level of 0.05, we reject the null hypothesis and accept the alternative hypothesis (*The accuracy of the random selection is lower than the accuracy of the machine selection*).

Using the original accuracy values (without rounding) from each machine and the mean value of the random selection for each participant presented in Table II, a normality test was performed, which indicated that the values follow the normal distribution. As a result, we applied the Student's t-test means for the machine hypothesis test, as presented in Table III.

When observing the values obtained for the p-value of

TABLE III
COMPARISON BETWEEN MACHINE LEARNING AND RANDOM SELECTION ACCURACIES. FIRST COLUMN HAS THE LEARNING MACHINE IDENTIFICATION, SECOND COLUMN CONTAINS THE P-VALUE OBTAINED BY THE T-STUDENT TEST. MEAN ACCURACIES ($\overline{Acc.}$) AND CONFIDENCE INTERVALS (CONF. INT.) FOR BOTH RANDOM SELECTION AND MACHINE LEARNING ARE GIVEN IN THE REMAINING COLUMNS.

		Random		Machine Learning	
	p.value	$\overline{Acc.}$	Conf. Int.	$\overline{Acc.}$	Conf. Int.
KNN	0.002	0.25	[0.25; 0.25]	0.33	[0.28; 0.38]
SVM	0.000	0.25	[0.25; 0.25]	0.37	[0.32; 0.41]
RF	0.000	0.25	[0.25; 0.25]	0.37	[0.32; 0.43]
NNET	0.000	0.25	[0.25; 0.25]	0.36	[0.31; 0.40]
LB	0.000	0.25	[0.25; 0.25]	0.35	[0.30; 0.40]

the Student's t-test from Table III, we can reject the null hypothesis and accept the alternative hypothesis, *The accuracy of the random selection is smaller than the accuracy of the selection via machine learning*. Thus, we can affirm that, in most cases, facial expression components (Action Units - AUs) can be employed to classify emotional categories self-reported during exposition to multimedia contents.

In a second experiment, all shots from all videos were grouped, and 60% of shots were randomly selected for training and the remaining (40%) for test. The employed machine learning were the ones that presented best results (in terms of Mean accuracies - $\overline{Acc.}$ and confidence intervals - Conf. Int.) in the previous experiment: SVM and Random Forest. This process was repeated 30 times to avoid any bias. The minimum and maximum values of the accuracies obtained for each participant are presented in the Table IV. The smallest participant's accuracies are highlighted in each row.

It can be seen from the Table IV that even the smallest accuracies are still higher than those shown in Table II. In this second evaluation, data from different frames of the same video could be part of both the training and test sets, which creates a better statistical distribution for classifier decision. Note that training and test data are distinct, but more strongly related since they may come from the same video exhibition. We believe that increasing the number of videos in the dataset may create a more representative data sample for training the models and thus improving the results of Table II.

The results presented on Tables II and IV were obtained using all available AU features as input. Next, we investigate the use of Boruta [18] feature selector to reduce the number of AUs as input to the learning models. The following relevance scores were assigned to a given feature, as a result of the feature selector decision: 1 if a feature is "Confirmed", 0.5 if it is labeled as "Tentative", and 0 otherwise. After 30 executions of the Boruta method, the average relevance score of each feature was calculated.

In order to check whether reducing the number of features would improve the classifiers, we selected the most relevant features per participant. To be considered relevant, we defined that the average feature relevance score had to be greater than or equal to 0.75. Table V shows the relevant features per participant. AUs 4, 6, 7, 10, 12, 14, 17, 25 and 26 were considered to be the most relevant features for more than half

TABLE IV
MINIMUM AND MAXIMUM ACCURACIES FOR RANDOM FOREST AND SVM MACHINE LEARNING WITH TRAINING AND TESTING USING 60% AND 40% PARTITIONING OF THE DATASET, RESPECTIVELY

Participant	Min.		Max.	
	Random Forest	SVM	Random Forest	SVM
01	0.750	**0.714**	0.920	0.857
02	0.663	**0.564**	0.832	0.752
03	0.344	**0.302**	0.531	0.510
04	0.719	**0.516**	0.859	0.719
05	0.781	**0.646**	0.896	0.844
06	0.757	**0.701**	0.875	0.840
07	0.602	**0.407**	0.707	0.593
08	0.786	**0.753**	0.916	0.909
09	0.855	**0.735**	0.949	0.872
10	0.722	**0.629**	0.844	0.775
11	0.747	**0.714**	0.844	0.851
12	0.661	**0.612**	0.843	0.826
13	0.746	**0.599**	0.894	0.775
14	0.492	**0.467**	0.639	0.672
15	0.767	**0.664**	0.905	0.845
16	0.346	**0.318**	0.514	0.439
17	0.667	**0.598**	0.780	0.735
18	0.415	**0.341**	0.585	0.528
19	0.762	**0.683**	0.884	0.835
20	0.478	**0.433**	0.627	0.590
21	0.630	**0.563**	0.785	0.726
22	0.636	**0.614**	0.750	0.758
23	0.764	**0.655**	0.885	0.811
24	0.662	**0.590**	0.799	0.719
25	0.647	**0.618**	0.824	0.779
26	0.678	**0.653**	0.810	0.843
27	0.779	**0.566**	0.882	0.750
28	0.659	**0.535**	0.791	0.682
29	0.808	**0.733**	0.933	0.933
30	0.588	**0.510**	0.719	0.699
31	0.611	**0.595**	0.825	0.762
32	0.711	**0.648**	0.852	0.796
33	0.623	**0.585**	0.764	0.792

of the participants by Boruta's algorithm.

All features labeled as relevant to each participant were used in a new experiment using the Random Forest model. Table VI shows accuracies with and without feature selection. The table also shows, out of all 17 features available, how many features were kept for the training after the feature selection. The highlighted accuracies are the lowest ones.

In order to verify the question "a reduced subset of AUs can produce better classification results than those produced when using all AUs", we proposed the following null hypothesis: *Employing all AUs in the experiments produce a accuracy that is greater than or equal to the accuracy obtained when employing the reduced subset of AUs*. To test this hypothesis, using the original accuracy values (without rounding) presented in Table VI for each participant, a normality test was performed, which indicated that those values follow the normal distribution. Then, we administered the Student's t-test, which obtained, for p-value = 0.023, mean accuracies and confidence intervals of 0.40 ([0.35;0.44]) and 0.38 ([0.32;0.43]), for the two types of experiments per participant, with reduced subset of AUs and with all AUs, respectively. From these results, we can reject the null hypothesis and accept the alternative hypothesis. Thus, the test indicates that in most cases "it

TABLE V
FEATURES THAT THE AVERAGE FEATURE RELEVANCE SCORE HAD TO BE GREATER THAN OR EQUAL TO 0.75

Participant	AU01	AU02	AU04	AU05	AU06	AU07	AU09	AU10	AU12	AU14	AU15	AU17	AU20	AU23	AU25	AU26	AU45
01	X	-	X	-	X	X	-	X	X	X	-	X	-	-	-	-	X
02	-	-	X	-	X	X	-	-	-	X	-	-	-	-	-	-	-
03	-	-	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-
04	X	X	X	X	X	X	X	X	X	X	-	X	X	X	-	X	-
05	X	X	X	-	X	X	X	X	X	X	X	X	X	-	X	X	X
06	-	-	X	-	X	X	-	X	X	X	X	X	-	-	-	X	-
07	-	-	X	-	X	X	-	X	X	X	-	-	-	-	X	-	-
08	X	X	X	-	X	X	X	X	X	X	X	X	X	X	X	X	X
09	X	X	X	X	X	-	X	X	-	X	X	X	-	X	X	X	-
10	-	-	X	X	X	X	X	X	X	X	-	X	-	X	-	X	X
11	X	X	X	-	X	X	-	X	X	-	-	X	-	X	-	X	-
12	-	-	X	X	X	X	X	X	X	X	-	-	X	-	X	X	-
13	X	-	X	X	X	X	-	X	X	X	-	X	X	X	X	X	X
14	X	-	-	-	X	X	-	X	X	X	X	X	X	-	X	-	-
15	X	-	X	-	X	X	-	X	X	X	-	-	-	X	X	X	X
16	-	-	-	-	X	-	-	-	-	-	-	-	-	-	-	-	-
17	-	-	X	X	X	X	-	X	X	X	X	-	-	X	-	X	-
18	-	-	X	-	-	-	-	-	-	X	-	-	-	-	X	-	-
19	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
20	-	-	X	-	X	X	X	X	X	X	-	-	-	X	-	-	-
21	X	X	X	-	X	X	-	X	X	X	X	X	-	-	X	X	-
22	-	-	X	-	X	-	-	X	X	X	X	X	X	-	X	X	-
23	X	X	X	-	X	X	-	X	X	X	X	X	-	X	X	-	X
24	X	-	X	X	X	X	-	-	-	X	-	X	-	-	X	X	X
25	X	X	X	-	X	X	-	X	X	X	-	X	X	-	X	-	X
26	-	-	X	X	X	X	-	X	X	X	-	X	-	-	X	-	-
27	-	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
28	X	X	X	X	X	X	X	-	X	-	X	X	-	X	X	X	X
29	-	-	X	-	X	X	-	X	X	X	-	-	-	-	X	X	-
30	-	-	X	-	X	-	-	X	X	X	-	X	-	-	X	-	-
31	-	-	X	-	X	X	X	X	X	X	-	X	X	X	X	X	-
32	X	X	X	-	X	X	X	X	X	X	X	-	X	X	X	-	X
33	-	X	X	-	X	X	-	X	X	X	X	-	-	-	X	X	X
Amount	16	13	31	11	31	27	12	27	27	29	14	21	12	15	23	20	14

is possible to find a reduced subset of AUs that produce classification results that are higher than those produced when using all AUs".

V. CONCLUSIONS

This paper shed light on the link between changes in human faces, while experiencing the display of a video, and the felt emotions. In order to facilitate the self-report of emotions, only 4 labels have been used, corresponding to the four quadrants of Russel's circumplex emotion model. To better answer the research question, a series of three experiments were carried out using a dataset of 20 videos of various genres.

In the first experiment, we investigated the use of five machine learning models to map Action Unit (AU) features into emotion labels. Training and test sets were created using data samples coming from distinct videos. This is probably a learning problem that is more challenging when compared to the case of using samples from the same videos. Similarity between frames is greatly decreased if choosing frames from different videos, which may lead to facial expressions that are more distinct. Most of the accuracies were above those of a random categorization (Table II).

In the second experiment, a 30-fold 60-40% (training/test) cross validation was performed, while allowing training/test sets contain data samples that, although distinct, could come from a same video. This time, accuracies were much better (Table IV), which may indicate that more video data is needed in order to better characterize the emotion patterns to be learned in the experiment reported on Table II. The results obtained by Table IV follow similar works such as Wang and Cheong [22] that used audiovisual features to rate scenes in 36 Hollywood films within 7 emotional labels, and Brezeale and Cook [23], who rated 81 videos in genres with closed captions and discrete cosine transform coefficients.

The third experiment provided some empirical evidence that the reduction in the number of features using the Boruta method produced better accuracy for most participants.

In this study, differently from previous related work, where the entire video is considered for labeling purposes (e.g. [23], [24], [25], [26], [27] and [28]), an emotional labeling investigation was presented by taking into account videos segmented into shots, which leads to more refined conclusions and applications of the results.

As future work, we propose to carry out new experiments after expanding the video dataset and participants, as well as adding other physiological data (e.g. heart rate, skin conductance), besides facial expressions, to evaluate the improvements in the accuracy of the learned models.

TABLE VI
COMPARISON RANDOM FOREST ACCURACIES FOR TRAINING AND TESTING WITH AND WITHOUT RELEVANT FEATURES SELECTION (FS) PER PARTICIPANT. THE ACCURACIES PRESENTED IN THE LAST COLUMN WERE OBTAINED USING ALL AVAILABLE AUS (17)

Participant	reduced features	with FS	without FS
01	9	0.383	**0.243**
02	4	**0.495**	0.505
03	1	0.250	**0.181**
04	14	0.342	**0.333**
06	9	0.160	**0.136**
07	7	0.464	**0.435**
10	12	**0.482**	0.506
11	10	0.269	**0.170**
13	14	**0.563**	0.632
14	10	0.438	**0.415**
15	11	0.500	**0.377**
16	1	**0.243**	0.252
17	10	0.256	0.256
18	3	0.261	**0.243**
20	8	0.339	**0.287**
21	12	**0.570**	0.579
22	10	**0.394**	0.417
23	13	0.428	0.428
24	10	**0.480**	0.504
25	12	0.478	**0.420**
27	16	**0.378**	0.385
28	14	0.549	**0.529**
29	8	0.417	**0.398**
30	7	0.225	0.225
31	12	**0.459**	0.484
32	14	0.580	**0.573**
33	11	0.328	**0.312**

Regarding facial expressions, we also intend to analyze the relationship between specific AUs and the accuracies obtained. Last and most importantly, we intend to investigate ways to use emotional state prediction based on facial expressions to devise novel personalized multimedia retrieval and summarization approaches.

REFERENCES

[1] H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg, "Movie genre classification via scene categorization," in *Proceedings of the international Conference on Multimedia - MM'10*. ACM Press, 2010.

[2] Y. Baveye, C. Chamaret, E. Dellandrea, and L. Chen, "Affective video content analysis: A multidisciplinary insight," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 396–409, oct 2018.

[3] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, may 1992.

[4] A. Hanjalic, "Extracting moods from pictures and sounds: towards truly personalized TV," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, mar 2006.

[5] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 410–430, oct 2015.

[6] E. A. Celik, "Affective analysis of videos: Detecting emotional content in real-life scenarios," Master's thesis, Technische Universitat Berlin, 2017.

[8] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, feb 2005.

[7] L. Canini, S. Benini, and R. Leonardi, "Affective analysis on patterns of shot types in movies," in *Proceedings of the 7th international symposium on Image and Signal Processing and Analysis (ISPA)*. Dubrovnik, Croatia: IEEE, 2011.

[9] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses," in *2008 Tenth IEEE International Symposium on Multimedia*. IEEE, dec 2008.

[10] A. Hanjalic, *Content-Based Analysis of Digital Video*. Springer, Berlin, 2013.

[11] J. A. Russell, "A circumplex model of affect." *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[12] J. A. Russell, M. Lewicka, and T. Niit, "A cross-cultural study of a circumplex model of affect." *Journal of Personality and Social Psychology*, vol. 57, no. 5, pp. 848–856, 1989.

[13] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proceeding of the 16th ACM international conference on Multimedia - MM 08*. ACM Press, 2008.

[14] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, may 2018.

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, sep 2010.

[16] T. Baltrusaitis, M. Mahmoud, and P. Robinson, "Cross-dataset learning and person-specific normalisation for automatic action unit detection," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, may 2015.

[17] A. Hildebrandt, S. Olderbak, and O. Wilhelm, "Facial emotion expression, individual differences in," in *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, 2015, pp. 667–675.

[18] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002. [Online]. Available: http://CRAN.R-project.org/doc/Rnews/

[19] M. Kuhn, "Building predictive models in r using the caret package," *Journal of Statistical Software*, vol. 28, no. 5, 2008.

[20] M. B. Kursa and W. R. Rudnicki, "Feature selection with theBoruta-Package," *Journal of Statistical Software*, vol. 36, no. 11, 2010.

[21] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta – a system for feature selection," *Fundamenta Informaticae*, vol. 101, no. 4, pp. 271–285, 2010.

[22] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 6, pp. 689–704, jun 2006.

[23] D. Brezeale, "Using closed captions and visual features to classify movies by genre," in *In Poster session of the Seventh International Workshop on Multimedia Data Mining (MDM/KDD2006*, 2006.

[24] S. Fischer, R. Lienhart, and W. Effelsberg, "Automatic recognition of film genres," in *Proceedings of the third ACM international conference on Multimedia - MULTIMEDIA'95*. ACM Press, 1995.

[25] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 52–64, jan 2005.

[26] H.-Y. Huang, W.-S. Shih, and W.-H. Hsu, "A film classifier based on low-level visual features," in *2007 IEEE 9th Workshop on Multimedia Signal Processing*. IEEE, 2007.

[27] Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li, "YouTubeCat: Learning to categorize wild web videos," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2010.

[28] D. Hazer, X. Ma, S. Rukavina, S. Gruss, S. Walter, and H. C. Traue, "Emotion elicitation using film clips: Effect of age groups on movie choice and emotion rating," in *Comm. in Computer and Information Science*. Springer International Publishing, 2015, pp. 110–116.