UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Operadores de fusão prévia para segmentação temporal de vídeo em cenas

Antonio Alessandro Rocha Beserra

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

ICMC USP
SÃO CARLOS

Antonio Alessandro Rocha Beserra

Operadores de fusão prévia para segmentação temporal de vídeo em cenas

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

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Rudinei Goularte

USP – São Carlos
Janeiro de 2023

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

B554o

Beserra, Antonio Alessandro Rocha
Operadores de fusão prévia para segmentação temporal de vídeo em cenas / Antonio Alessandro Rocha Beserra; orientador Rudinei Goularte. -- São Carlos, 2023.
78 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Segmentação de Vídeo. 2. Fusão Prévia. I. Goularte, Rudinei, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Antonio Alessandro Rocha Beserra

Early fusion operators for temporal video scene segmentation

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Rudinei Goularte

USP – São Carlos
January 2023

AGRADECIMENTOS

Agradeço a todos os colegas do laboratório Intermídia, especialmente ao Rodrigo, Tiago e Marcus, por todo o apoio prestado de suma relevância para a condução deste mestrado.

Agradeço ao meu orientador, Rudinei, pela oportunidade e apoio dados durante a realização deste mestrado, assim como pelos momentos de paciência e compreensão.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"O opressor não seria tão forte
se não tivesse cúmplices entre os próprios oprimidos."
(Simone de Beauvoir)

RESUMO

BESERRA, A. A. R. **Operadores de fusão prévia para segmentação temporal de vídeo em cenas**. 2023. 78 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Técnicas de fusão prévia têm sido propostas em tarefas de análise multimídia como uma maneira de melhorar a eficácia ao gerar representações de dados mais compactas, expressivas e capazes de preservar a semântica presente nos dados. Os trabalhos recentes no domínio de vídeo digital empregam multimodalidade fazendo jus à natureza multimodal de um vídeo. Esse espaço heterogêneo, somado à dificuldade de se obter uma etapa de fusão prévia desacoplada e separável do restante do processamento, limita possíveis melhorias que poderiam ser alcançadas nas etapas isoladamente. Além disso, técnicas foram projetadas para problemas específicos, não podendo ser generalizadas, o que também as tornam inseparáveis da tarefa de análise de vídeo em questão. Motivado por esse cenário, este trabalho de mestrado propõe a aplicação dos operadores de fusão prévia, Soma, Máximo e Concatenação, que atuem no médio nível semântico, desacoplando o operador de qualquer tarefa específica e, ao mesmo tempo, com um custo computacional mais simples. Os operadores foram aplicados em duas bases de dados publicamente disponíveis da tarefa de Segmentação Temporal de Vídeo em Cenas. Os resultados atingidos competem com os do estado da arte com a vantagem de simplicidade computacional.

Palavras-chave: Segmentação de Vídeo, Fusão Multimodal, Fusão Prévia.

ABSTRACT

BESERRA, A. A. R. **Early fusion operators for temporal video scene segmentation**. 2023. 78 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Early fusion techniques have been proposed in multimedia analysis tasks as a way to improve efficiency by generating more compact, expressive data representations capable of preserving the semantics present in the data. Recent work in the digital video domain employs multimodality due to the multimodal nature of a video. This heterogeneous space, added to the difficulty of obtaining a early fusion step that is uncoupled and separable from the rest of the processing, limits possible improvements that could be achieved in the steps isolatedly. Furthermore, techniques were designed for specific problems and cannot be generalized, which also makes them inseparable from the video analysis task. Motivated by this scenario, this master's degree work proposes the application of the early fusion operators, Sum, Maximum and Concatenation, which act at the medium semantic level, decoupling the operator from any specific task and, at the same time, with a simpler computational cost. The operators were applied in two publicly available databases of the Temporal Video Scene Segmentation task. The achieved results competed with those of the state of the art with the advantage of computational simplicity.

Keywords: Video Segmentation, Multimodal Fusion, Early Fusion.

LISTA DE ILUSTRAÇÕES

LISTA DE ALGORITMOS

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

1.1 Contextualização

Cada vez mais as pessoas interagem com vídeo digital, seja consumindo-o ou produzindo-o. Para as mais diversas finalidades, como se entreter, estudar ou acompanhar as notícias, vídeo tem sido assistido por meio de variados dispositivos, tais como televisões, smartphones, tablets e até mesmo em centrais multimídia veiculares. Mais de 2 bilhões de usuários, quase um terço dos usuários da Internet, corresponde ao número de usuários no Youtube[1]. Enquanto mais de 500 horas de vídeo são enviadas para os seus servidores a cada minuto por parte desse público, mais de 700 mil horas são assistidas na plataforma (Youtube, 2021). Uma das maiores plataformas de streaming de vídeo, Netflix[2], reportou ter mais de 220 milhões de assinantes no fim do terceiro trimestre de 2022 (Netflix, 2022). Além dessas estatísticas, o uso crescente de aplicativos de vídeo, tais como TikTok e Kwai, mostra que o tráfego na Internet correspondente ao acesso multimídia tem crescido nos últimos anos e que essa tendência continuará (CETIC, 2021).

A popularização desse tipo de mídia se deu devido à melhoria das taxas de transmissão pelas redes e mediante a facilidade e rápida difusão de meios de se produzir e consumir vídeo digital (HO; LO; FENG, 2008). Aparelhos móveis, como smartphones e tablets, permitiram não só reproduzir esse tipo de conteúdo, como também gravá-lo, armazená-lo, editá-lo e transmiti-lo. As crescentes disponibilidade e velocidade de redes, especialmente as móveis, também contribuíram para popularizar esse acesso (WEI; BHANDARKAR; LI, 2007). Com isso, pesquisadores e mercado trabalham em soluções tecnológicas que tornam a experiência multimídia para o usuário final mais fácil e intuitiva (ABDU; YOUSEF; SALEM, 2021; SUMALAKSHMI; VASUKI, 2022; RAVAL; GOYANI, 2022).

A vasta quantidade de vídeo digital disponível online dificulta processamento relacio-

[1] <www.youtube.com>
[2] <www.netflix.com>

nado a tarefas computacionais, como a recuperação de conteúdo, por exemplo (MYLONAS; AVRITHIS, 2008). Essa dificuldade ocorre devido ao problema da Sobrecarga de Informação (GROSS, 1965), datado há décadas, mas que é enfatizado atualmente no meio digital. Em termos gerais, esse problema retrata a situação que uma pessoa enfrenta ao ter que lidar com uma grande quantidade de dados para encontrar uma informação específica que lhe interessa em um tempo hábil. Essa situação é melhor percebida no meio online, por meio de posts em redes sociais, e-mails e artigos científicos, por exemplo, conteúdos esses que são publicados ou enviados em grande quantidade diariamente. No domínio de vídeo digital, esse problema pode ser sentido por um usuário de uma plataforma de streaming de vídeos B2C (*Business to Consumer*), como Youtube ou Dailymotion[3]. A título de exemplificação, um usuário gostaria de ver um vídeo, já assistido no passado, sobre um produto que ele almeja comprar. Mesmo que ele se lembre da string de busca usada para encontrar esse vídeo, encontrá-lo novamente pode se tornar uma tarefa difícil devido a novos vídeos sobre esse produto aparecerem nos resultados da busca desde a última consulta.

Os avanços nas soluções tecnológicas para as necessidades e problemas de sistemas multimídia, tais como os decorrentes da Sobrecarga de Informação (ALAM; ULLAH; LEE, 2020), estabelecem uma área de pesquisa: Análise Multimídia (CHINCHOR; CHRISTEL; RIBARSKY, 2010; POUYANFAR *et al.*, 2018). Essa área abrange diversas tarefas que envolvem processamento e análise de dados multimídia. Dentre essas tarefas podem ser citadas: sumarização, recomendação e personalização de conteúdo multimídia. A primeira consiste em sintetizar um conteúdo de modo a preservar suas partes mais relevantes; a segunda, em sugerir conteúdo de interesse ao usuário sem uma demanda explícita; e a terceira, em estruturar e apresentar uma informação de tal maneira que melhor satisfaça as necessidades do usuário automaticamente (FOSS *et al.*, 2019).

Trabalhos iniciais em Análise Multimídia processavam apenas o canal de informação visual do vídeo. Além desse canal, também chamado modalidade, um vídeo também pode ser composto de canais de informação aurais e textuais (SNOEK; WORRING, 2002). Com o surgimento de técnicas multimodais nas últimas décadas, modalidades passaram a ser combinadas, resultando em maior ganho de eficácia (ATREY *et al.*, 2010). Sidiropoulos *et al.* (2011) e Lopes, Trojahn e Goularte (2014) enfatizaram que técnicas unimodais se restringem a domínios específicos e que mais de uma modalidade deve ser levada em consideração para se obter melhores resultados. Assim, tarefas de análise de vídeo digital requerem processar múltiplas modalidades devido a sua intrínseca natureza multimodal.

Métodos multimodais empregam um processo de fusão multimodal, no qual dados de diferentes modalidades são combinados (ATREY *et al.*, 2010). Essa fusão pode ser classificada como: fusão prévia, na qual os dados de entrada são combinados antes da etapa de decisão; fusão tardia, na qual os dados de cada modalidade passam pela etapa de decisão separadamente e então

[3] <www.dailymotion.com>

essas decisões são fundidas em uma única consensual, e a fusão híbrida, que consiste na fusão prévia de algumas modalidades assim como na fusão tardia de algumas decisões. A etapa de decisão aqui referida diz respeito ao domínio da aplicação.

A fusão de informação torna-se necessária em contextos multimodais (ATREY *et al.*, 2010). Ela encontra uma representação única dos dados oriundos seja de diversas modalidades ou das diversas decisões em uma tarefa. Ao mesmo tempo em que essa representação é mais compacta, ela ainda deve ser representativa de forma a expressar a semântica relevante contida nos dados. Se a correlação entre características de diversas modalidades for devidamente preservada, melhores resultados nas tarefas de análise multimídia podem ser alcançados (TROJAHN; KISHI; GOULARTE, 2018). Esse tipo de correlação costuma ser ignorado na fusão tardia. Além disso, se o processo de fusão prévia for menos custoso do que realizar n vezes a etapa de decisão para as n modalidades ou características envolvidas, a fusão prévia também se torna mais vantajosa em relação ao custo computacional (KISHI; TROJAHN; GOULARTE, 2018).

As primeiras tentativas em fusão prévia a fizeram de uma forma ingênua, como concatenar representações de dados de baixo nível semântico (PEREIRA JR.; FERRAZ; GONZAGA, 2018). Estudos como os de Wang *et al.* (2017) e Pereira Jr., Ferraz e Gonzaga (2018) contornaram isso ao fundir esses vetores com operações matemáticas como Soma e Gram (PARK *et al.*, 2018). Contudo, a fusão de vetores de características de diversos descritores em um espaço heterogêneo e a falta de qualquer correlação intermodal ainda deixou espaço para melhorias de eficácia nas tarefas de análise multimídia (JHUO *et al.*, 2014; KISHI; TROJAHN; GOULARTE, 2019).

As operações de fusão prévia com operadores mais simples no baixo nível semântico foram superadas quando se utilizaram técnicas, consideravelmente mais complexas, que fundiam características de médio nível semântico (GAONKAR *et al.*, 2021). Essas técnicas combinavam análise para detectar e representar informação proveniente de diferentes modalidades, fusão de informação e procedimentos orientados a tarefas. Isso culminou em representações mais expressivas, auxiliando tarefas de análise de vídeo, como a Segmentação Temporal de Vídeo em Cenas, a atingirem maior eficácia. O interesse nessas técnicas ofuscou a possibilidade de empregar operadores de fusão mais simples no espaço de características de médio nível semântico, isolado de qualquer processo computacional complexo, em que as desvantagens oriundas da fusão de informação no baixo nível são reduzidas.

1.2 Problema

A fusão prévia se depara com problemas como a heterogeneidade dos dados, dimensionalidades diferentes e a falta de sincronização entre informações de modalidades diferentes. Tarefas de análise de vídeo que empregam fusão prévia são projetadas como um *framework* monolítico, tornando as etapas inseparáveis e adicionando complexidade computacional ao processo. A fusão de informação evoluiu de operadores simples para operações mais complexas, aumentando

a expressividade da representação fundida. Contudo, isso tornou a etapa de fusão altamente acoplada ao processo como um todo, deixando-a dependente da tarefa.

1.3 Objetivo

O objetivo principal deste trabalho consiste em propor a aplicação de operadores de fusão prévia multimodal para características de médio nível semântico em tarefas de análise de vídeo. Ao realizar a fusão com operadores de fusão prévia em um espaço de características com maior nível semântico, espera-se que as representações obtidas não acarretem perda significante de eficácia quando aplicadas às tarefas em questão comparado a técnicas do estado da arte. Com o uso de operadores espera-se também um ganho no processamento devido a sua inerente simplicidade computacional.

1.4 Organização

Esta dissertação está organizada nos seguintes capítulos. No Capítulo 2, conceitos necessários para a compreensão deste trabalho são apresentados, tais como hierarquia estrutural de vídeo digital, extração de características nos diferentes níveis, multimodalidade e tipos de fusão. O Capítulo 3 descreve as lacunas presentes em técnicas de fusão prévia multimodal, que são acopladas a tarefas de análise de vídeo. O Capítulo 4 detalha os operadores de fusão prévia que foram aplicados. O Capítulo 5 relata os experimentos realizados para avaliar a aplicação desses operadores. O Capítulo 6 apresenta as conclusões resultantes do trabalho, bem como limitações do trabalho realizado, sugestões de trabalhos futuros e as publicações originadas desta pesquisa.

CAPÍTULO

2

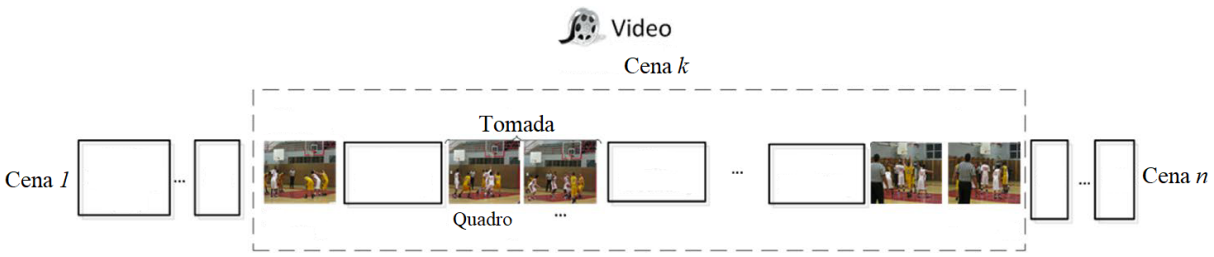
FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é apresentada a fundamentação teórica com os conceitos necessários para a devida compreensão deste trabalho. Na Seção 2.1, o conceito de vídeo digital e sua estrutura hierárquica são abordados. Na Seção 2.2, a definição de característica (do Inglês: *feature*) e seus tipos são discutidos. Na Seção 2.3, multimodalidade e estratégias de fusão multimodal são apresentadas. Por fim, bases de dados confiáveis e medidas de avaliação usadas em tarefas de análise de vídeo são discorridas na Seção 2.4.

2.1 Vídeo Digital

Vídeo pode ser definido como uma sequência sucessiva de imagens dispostas temporalmente em uma frequência suficiente para causar um estímulo visual contínuo ao espectador, podendo ou não conter uma trilha de áudio (CHAPMAN; CHAPMAN, 2009; HAVALDAR; MEDIONI, 2009). Cada uma dessas imagens é denominada de quadro e a frequência em que eles são exibidos é medida em quadros por segundo ou FPS (*Frames Per Second*, do inglês). A Figura 1 exibe a estrutura hierárquica de vídeo composta por quadros, tomadas e cenas.

Figura 1 – Estrutura hierárquica de vídeo



Fonte: Adaptada de Güder e Çiçekli (2017).

Uma tomada é uma sequência de quadros que foi obtida ininterruptamente por uma única câmera (ZHANG; LOW; SMOLIAR, 1995). Cena é uma sequência de tomadas adjacentes semanticamente relacionadas (SARACENO; LEONARDI, 1997). Embora o conceito de cena tenha sido definido de diversos modos por diversos pesquisadores (RUI; HUANG; MEHROTRA, 1999; HANJALIC; LAGENDIJK; BIEMOND, 1999; SUNDARAM; CHANG, 2002; COUR *et al.*, 2008), esse é o mais aceito e empregado no que concerne os interesses de usuários de sistemas multimídia, além de ser também independente do domínio da aplicação. Segmentar um vídeo em quadros é trivial por sua natureza. Segmentá-lo em tomadas já é considerado um problema essencialmente resolvido (KRAAIJ; SMEATON; OVER, 2004). Contudo, a segmentação temporal de um vídeo em cenas ainda é um problema em aberto (FABRO; BÖSZÖRMENYI, 2013).

A principal dificuldade inerente ao uso da definição de cena de Saraceno e Leonardi (1997) é conceber um algoritmo que seja capaz de mensurar a relação semântica entre duas tomadas. Conceitos semânticos de alto nível, que podem ser usados para medir a coesividade entre tomadas, não são facilmente interpretáveis por computadores. Como exemplo, uma pessoa assistindo a um vídeo poderia identificar objetos, pessoas e suas características tais como um carro azul e seu modelo ou um cantor pela sua música que passa na trilha de áudio de um vídeo. Como programar um computador para extrair esse tipo de informação de um vídeo, que basicamente consiste na sucessão de matrizes de valores de pixels, é um problema retratado por Smeulders *et al.* (2000) como Lacuna Semântica. Essa lacuna refere-se à distância entre a representação computacional de um dado e o seu significado quando interpretado pelo usuário. Como uma maneira de atenuá-la, existem técnicas para se extrair características a partir dos dados brutos a fim de auxiliar a inferir a semântica contida neles (SNOEK; WORRING, 2009; MACFARLANE, 2016).

2.2 Características de Vídeo

Sistemas multimídia lidam com objetos multimídia tais como imagens e sons. Ações como a comparação entre dois objetos tornam-se intratáveis a partir de suas representações de dados brutos devido aos problemas da lacuna semântica e sobrecarga de informação, já citados anteriormente. Ao invés disso, essas ações são realizadas com representações alternativas que caracterizam algum aspecto desse conteúdo original. Esses aspectos constituem propriedades distinguíveis desses objetos e são denominadas de características (ATREY *et al.*, 2010). As representações dessas características possuem menor volume de dados que o conteúdo original, são normalmente codificadas em um vetor (também chamado de vetor de características) e podem ser classificadas de acordo com o nível semântico representado por elas: baixo, médio e alto (MARTINET; SAYAD, 2012). Elas também podem evidenciar informação que antes não estava explicitamente codificada nos dados originais. Em uma visão geral, vetores de características são uma representação compacta de uma mídia usados para comparar mídias entre si.

2.2.1 Características de Baixo Nível

Descritores de características de baixo nível são medidas numéricas extraídas diretamente do dado bruto sem envolver um processo de aprendizado de máquina ou análise estatística de outros documentos do mesmo tipo (MARTINET; SAYAD, 2012). Vídeos podem ser representados por características extraídas de cada uma de suas modalidades: visual, aural e textual. Um programa que calcula essas características de uma mídia é conhecido como detector ou extrator (TUYTELAARS; MIKOLAJCZYK, 2008).

2.2.1.1 Características Visuais

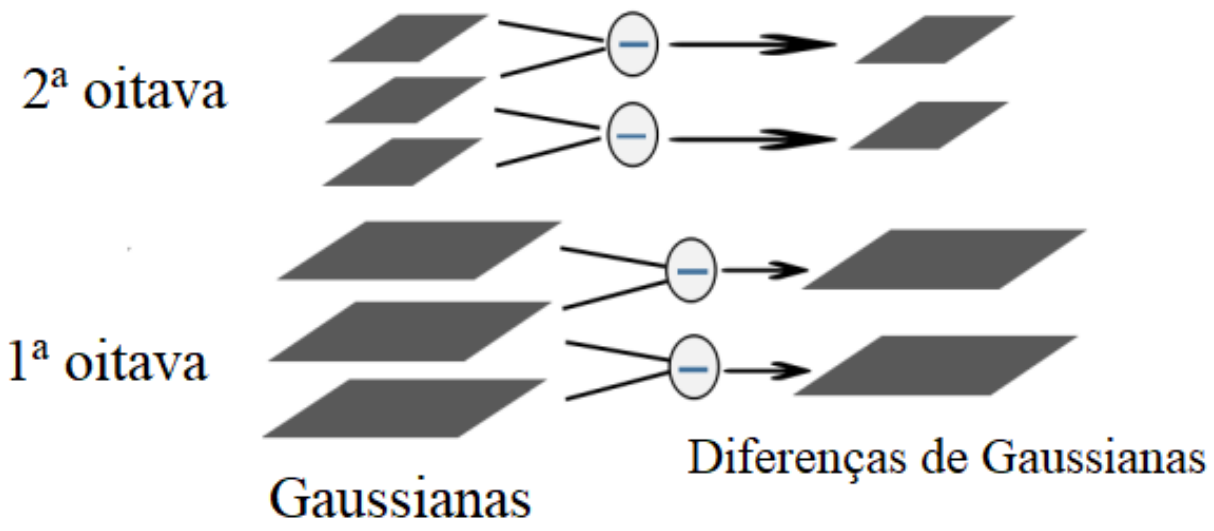
Características visuais podem ser divididas de acordo com seu escopo de representação: global e local. Características globais representam a imagem como um todo, já as locais, apenas algumas regiões específicas da mesma.

Uma característica visual global bastante usada para comparação de imagens é o histograma de cor (NAGASAKA; TANAKA, 1992; SUHASINI; KRISHNA; KRISHNA, 2016). Esse descritor corresponde à frequência em que as cores de uma imagem aparecem. Pode ser implementado como um vetor em que cada índice corresponde a um intervalo de cores e o valor naquele índice representa o número de pixels da imagem cuja cor se situa no intervalo associado àquele índice. A computação dessa característica é de baixa complexidade, além de os histogramas serem invariantes a operações de escala e rotação. Contudo, como os histogramas de cor não levam em consideração a informação espacial de um pixel, eles falham quando operam em imagens diferentes mas com distribuição de cores similares. Histogramas de cor estão presentes em diversos trabalhos de tarefas de análise de vídeo (KENDER; YEO, 1998; SIDIROPOULOS *et al.*, 2011; JI *et al.*, 2019).

Um descritor local corresponde a uma região de uma imagem que contém um padrão que difere de sua vizinhança (TUYTELAARS; MIKOLAJCZYK, 2008). Com isso, apenas padrões relevantes de uma imagem são codificados. Um bom extrator de características visuais deve garantir repetibilidade e precisão, sendo invariante a mudanças geométricas e fotométricas, ao mesmo tempo em que deve ser capaz de distinguir objetos diferentes (ABDEL-HAKIM; FARAG, 2006; GRAUMAN; LEIBE, 2011). Como exemplo desse tipo de descritor, podemos citar o *Scale-Invariant Feature Transform* (SIFT) (LOWE, 2004), usado em tarefas de análise multimídia (PATEL; DABHI; PRAJAPATI, 2020; MAHMOODZADEH, 2021; FOX; SCHOEFFMANN, 2022).

As características SIFT de uma imagem são geradas ao identificar máximos e mínimos locais em uma pirâmide de diferença de gaussianas em versões redimensionadas dessa imagem, como ilustrado na Figura 2. Para cada escala de redimensionamento, a versão da imagem original é incrementalmente convoluída com um filtro gaussiano. A cada um desses conjuntos dá-se o nome de oitava. Computa-se então as diferenças de gaussianas entre imagens consecutivas na

Figura 2 – Pirâmide de gaussianas e suas diferenças



Fonte: Adaptada de Li (2017).

mesma oitava. Os pontos máximos e mínimos locais são detectados comparando-se cada ponto com sua vizinhança 3x3 nas matrizes de diferenças em uma oitava. O ponto é selecionado como um candidato a ponto-chave se for maior ou menor que seus 26 vizinhos. Para determinar a localização interpolada dos pontos-chave, Brown e Lowe (2002) ajustaram uma função quadrática 3D por meio de expansão de Taylor. Após isso, pontos-chave com baixo contraste ou alta resposta de aresta em uma única direção são descartados. Para cada ponto são associadas uma magnitude de gradiente e uma orientação computadas usando diferenças de pixels em uma vizinhança de 4 pixels. O descritor SIFT de um ponto-chave é obtido ao se calcular histogramas de orientações de 8 posições em 16 subjanelas 4 x 4 de uma janela de 16 x 16. Esses histogramas concatenados formam o vetor de características SIFT. Esse vetor é normalizado para um comprimento unitário, limiarizado para 0.2 e renormalizado. Assim, o vetor de características SIFT tem 128 dimensões.

Apesar de as operações envolvidas no cálculo do vetor de características SIFT garantirem que ele é invariante a escala e rotação, cor não é levada em conta nesse descritor. Devido a isso, apenas cantos e junções T são evidenciados por ele, que são características indiferentes a cor. Isso pode ser exemplificado na Figura 3. Nas Figuras 3a e 3b são exibidos dois produtos diferentes que se diferenciam majoritariamente por sua cor e os pontos-chave SIFT extraídos a partir dessas imagens são destacados com círculos coloridos por meio da biblioteca OpenCV Python[1]. Uma demonstração da invariância do SIFT à cor pode ser vista na Figura 3c. Nela é apresentada uma correspondência entre os 32 pontos-chave mais similares das Figuras 3a e 3b, calculada com distância euclidiana entre os seus respectivos descritores. Observa-se que algumas regiões de contorno em volta da palavra que especifica o tipo do produto (capim-cidreira ou

[1] <https://pypi.org/project/opencv-python/>

camomila) foram consideradas similares, embora a cor da borda externa delas seja diferente.

Figura 3 – Exemplo de invariância do SIFT à cor

(a) Pontos-chave SIFT da imagem de um produto

(b) Pontos-chave SIFT da imagem de outro produto

(c) Os 32 pontos-chave SIFT mais similares entre os dois

Fonte: Elaborada pelo autor.

Abdel-Hakim e Farag (2006) propuseram uma variante do SIFT que leva em consideração a cor, chamada Colored SIFT (CSIFT). Eles o fizeram ao modelar o espaço de cores RGB em um Gaussiano e calculando as magnitudes de gradiente e histogramas assim como é feito no SIFT. Assim, o vetor de características CSIFT possui 3 x 128 = 384 dimensões. Com a cor sendo levada em consideração nesse processo, vetores de característica CSIFT não só evidenciam cantos e junções T, mas também as cores empregadas em suas regiões. Na Figura 4 o exemplo anterior agora é tratado com CSIFT ao invés de SIFT. Os pontos-chave CSIFT são exibidos nas Figuras 4a e 4b. Nota-se que os pontos-chave são praticamente os mesmos entre os dois extratores. A diferença pode ser notada nos descritores. Os 32 pontos-chave CSIFT mais similares

foram correspondidos assim como no exemplo anterior e essa associação é exibida na Figura 4c. Percebe-se que as correspondências entre as palavras "capim-cidreira" e "camomila" que antes constavam na comparação entre vetores SIFT não figuram mais neste caso. Numericamente falando, a diferença também é perceptível. Se antes a média das distâncias entre os vetores SIFT correspondentes aos 32 pontos-chave era de 35,46, essa média aumenta para 307,55 no caso de vetores CSIFT. Isso evidencia que CSIFT consegue distinguir imagens similares com cores diferentes melhor do que seu antecessor (LAKSHMI; VAITHIYANATHAN, 2017).

Figura 4 – Exemplo de extração de características CSIFT

(a) Pontos-chave CSIFT da imagem 3a

(b) Pontos-chave CSIFT da imagem 3b

(c) Os 32 pontos-chave CSIFT mais similares entre os dois

Fonte: Elaborada pelo autor.

2.2.1.2 Características Aurais

Características aurais representam aspectos e propriedades específicas de sinais de áudio em uma forma compacta. Essas características podem ser classificadas de acordo com o domínio

em que a natureza da fonte de informação extraída se situa (MITROVIĆ; ZEPPELZAUER; BREITENEDER, 2010; ALÍAS; SOCORÓ; SEVILLANO, 2016). Um descritor bastante utilizado e que reside no domínio cepstral é o *Mel-Frequency Cepstral Coefficients* (MFCC) (DAVIS; MERMELSTEIN, 1980).

MFCC foi originalmente proposto para reconhecimento de palavras monossílabas. Contudo, ele também demonstrou sua utilidade em reconhecimento de fala e classificação de conteúdo aural (JOTHILAKSHMI; GUDIVADA, 2016; ALÍAS; SOCORÓ; SEVILLANO, 2016). Ele se baseia na escala mel, uma unidade de medida baseada em como os humanos percebem frequências de som (STEVENS; VOLKMANN; NEWMAN, 1937). Devido a essa percepção não ser linear, essa escala é um espaço de frequência linear abaixo de 1 kHz e um logarítmico acima desse valor. Os passos para se obter os coeficientes cepstrais da frequência mel a partir de um sinal de áudio são (RAO; KOOLAGUDI, 2013):

1. Pré-enfatizar o sinal de áudio;
2. Dividi-lo em uma sequência de janelas, comumente com um tamanho de 20 ms e um deslocamento de 5 ms e aplicar um janelamento de Hamming em cada uma;
3. Computar a magnitude do espectro para cada janela ao aplicar a transformada Discreta de Fourier;
4. Ao passar o sinal, que está no domínio de frequência, através de um banco de filtros de Mel, o espectro Mel é obtido;
5. Para então obter os valores MFCCs, uma Transformada Discreta de Cosseno é aplicada ao logaritmo dos valores obtidos no passo anterior. Os treze primeiros coeficientes resultantes compõe o vetor de características MFCC para cada janela calculada.

Esses coeficientes cepstrais têm papel notório em reconhecimento de fala devido a sua capacidade em representar de forma compacta a informação relevante em uma janela de curto período de tempo de um sinal de áudio contínuo (MORGAN; BOURLARD, 1995). Desde então, MFCC é considerado estado da arte em aplicações envolvendo processamento de áudio, tais como recuperação, recomendação e reconhecimento (YANG *et al.*, 2019). Apesar de sua ampla utilização, ele ainda é passível de melhora, já que sua performance degrada consideravelmente na presença de ruído (SINGH; CHETTY, 2012; SHARMA; ALI, 2015). Mesmo assim, ele demonstra ser relevante em tarefas de análise de vídeo (IMRAN *et al.*, 2019; POORNA *et al.*, 2020; LIN *et al.*, 2021; SHARAFI *et al.*, 2022).

2.2.1.3 Características Textuais

O canal textual de um vídeo pode incluir dados como legendas. Texto pode ser extraído a partir dos outros canais com o uso de reconhecimento de fala ou caracteres, quando a legenda

de um vídeo é "gravada" na modalidade visual. Assim, texto acaba sendo usado como uma informação complementar a esses canais. Embora haja métodos que comparem dados textuais brutos sem qualquer processamento, também há extratores textuais mais elaborados, derivados das áreas de Processamento de Linguagem Natural e Recuperação de Informação. Um exemplo popular disso é o *Bag-of-Words* (BoW), que consiste numa representação compacta de um texto (AGGARWAL; ZHAI, 2012).

Em BoW, um texto é representado por um vetor de palavras quantificadas, de acordo com um dicionário computado. Como o vetor resultante pode ter alta dimensionalidade, técnicas para reduzi-la são empregadas tais como remoção de palavras vazias, que são palavras que não possuem muito sentido quando estão fora de um contexto, como artigos e preposições, e radicalização, que consiste em reduzir palavras morfologicamente similares (flexionadas ou derivadas) em um único fragmento, que é o seu radical (RAJARAMAN; ULLMAN, 2011). A partir de vetores BoW, dois documentos podem ser comparados por alguma função de similaridade entre eles, como a similaridade por cosseno, por exemplo. Essa técnica deu origem a versões alternativas que operam em conteúdos visuais e aurais resultando em um termo mais genérico, *Bag-of-Features* (BoF) (CSURKA *et al.*, 2004; MÜHLING *et al.*, 2012). Por auxiliarem a detecção de aspectos semânticos de modo latente e, consequentemente, reduzirem a lacuna semântica presente em vídeos, vetores BoF são classificados em um nível intermediário entre características de baixo e alto níveis.

2.2.2 Características de Médio Nível

Sistemas de indexação multimídia mapeiam correspondências entre características de baixo nível e alto nível. Uma vez que esse modelo é construído, o sistema consegue inferir conceitos semânticos a partir de características de baixo nível. Características de médio nível auxiliam nesse processo ao afunilar a lacuna que há entre esses dois níveis. Eles são obtidos após uma análise automática das características de baixo nível em uma coleção ou subconjunto de documentos (MARTINET; SAYAD, 2012). Um exemplo desse tipo de característica é o BoF, que é uma generalização do modelo BoW para outros tipos de dados além de texto, como imagem e áudio.

A primeira pesquisa com BoF veio da área de Visão Computacional substituindo documentos textuais por imagens e palavras por *textons*, vetores que caracterizam texturas (LEUNG; MALIK, 2001). Mais tarde, Csurka *et al.* (2004) propuseram a abordagem de *Bag-of-Keypoints* usando centroides de clusters de características de baixo nível, SIFT nesse caso, ao invés de *textons*. Tal abordagem ficou sendo conhecida e referenciada como *Bag-of-Visual-Words* (BoVW) em trabalhos futuros. Mühling *et al.* (2012) também propuseram um método nessa mesma linha usando vetores de características MFCC e o designaram *Bag-of-Auditory-Words* (BoAW). Apesar de haver variações no modelo BoF, duas mídias podem ser comparadas a partir de suas características de baixo nível extraídas pelo mesmo descritor. Esse processo é descrito no Algoritmo 1.

Algoritmo 1 – Comparação de mídias a partir de vetores BoF

Entrada: M_1, M_2: Mídias a serem comparadas; d: uma função de distância; E: um extrator de características
Retorno: D: a dissimilaridade entre as mídias

1: $v_1 \leftarrow E(M_1)$
2: $v_2 \leftarrow E(M_2)$
3: $V \leftarrow v_1 \cup v_2$
4: Disponha o conjunto V em um conjunto K de subconjuntos $k_1, k_2, \cdots, k_n$
5: $C \leftarrow \{\}$
6: **para** $k \in K$ **faça**
7: Calcule o descritor c que melhor representa o subconjunto k
8: $C \leftarrow C \cup c$
9: **fim para**
10: Compute H_1, o histograma de frequências de cada característica $c \in C$ em M_1
11: Compute H_2, para M_2
12: **retorna** $d(H_1, H_2)$

Os conjuntos de vetores de características de baixo nível extraídos por E das mídias M_1 e M_2 são representados por v_1 e v_2, respectivamente. Os subconjuntos $k_1, k_2, \ldots, k_n$ de vetores de características também são chamados de clusters ou bags. Para cada k_i é computado um vetor que melhor o caracteriza, c_i. Esse vetor pode ser o centroide ou medoide do grupo e também é chamado de característica, daí o nome *Bag-of-Features*. O conjunto C define o dicionário de características. Os histogramas BoF de cada documento são calculados a partir da contagem de ocorrências de cada característica no conjunto de vetores de características desse documento. O passo 4, por definir em quantos subconjuntos o conjunto V será particionado, também define o número de características e, consequentemente, o tamanho do dicionário. Por não ser uma etapa trivial, costumam ser empregadas técnicas de aprendizado de máquina para determinar o melhor tamanho de dicionário (CSURKA *et al.*, 2004; MÜHLING *et al.*, 2012).

2.2.3 Características de Alto Nível

Características de alto nível descrevem informações interpretáveis por humanos extraídas de uma mídia ou parte dela (MARTINET; SAYAD, 2012). Elas costumam ser expressadas como um conjunto de palavras chave ou uma descrição textual. Como exemplo, podemos citar os metadados e tags, embora nem todas as mídia contenham dados categorizados dessa forma (CAMBRIA; HUSSAIN, 2012).

Um exemplo de aplicação de características de alto nível em um sistema multimídia seria um usuário buscando por vídeos de análise sobre um produto que ele tenha interesse em comprar. Uma provável string de busca inserida por esse usuário seria "review" seguida do nome do produto. O sistema então poderia retornar vídeos em que aparecem outras pessoas fazendo

uma análise sobre esse produto. Embora atualmente hajam plataformas que dispõem desse tipo de consulta, como o Youtube, elas operam com dados anotados manualmente pelos próprios usuários como a descrição e tags do vídeo.

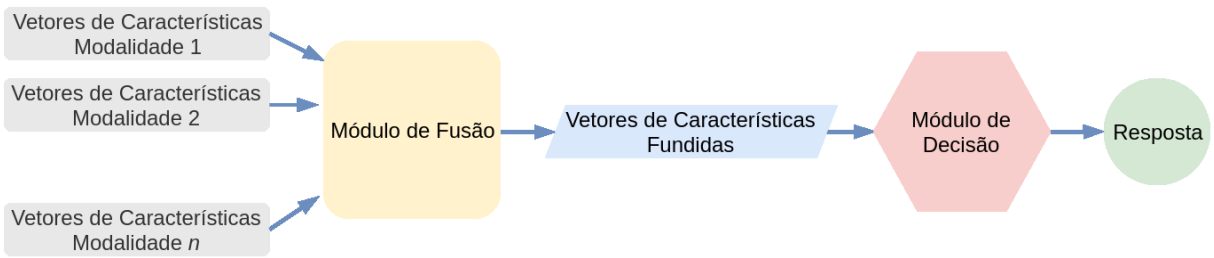
2.3 Multimodalidade

A concepção de multimodalidade é originado da teoria da comunicação (KRESS, 2005). Snoek e Worring (2002) estenderam o conceito de multimodalidade de Nigay e Coutaz (1993) para vídeo como a capacidade do criador de um vídeo de expressar uma ideia semântica predefinida usando pelo menos dois canais de informação. Esses canais de informação, ou modalidades, são o visual, aural e textual. Como a informação em vídeo é transmitida através de diferentes canais, a análise e processamento em apenas um deles pode não capturar todo o conceito semântico representado naquele contexto. Para conceitos que se expressam em mais de uma modalidade, exige-se o emprego de alguma estratégia que combine as informações provenientes de diferentes fontes.

Ao aplicar multimodalidade em um sistema multimídia faz-se necessária a definição de uma estratégia para combinar dados de diferentes canais de comunicação. Com isso, pode-se aproveitar e evidenciar a correlação entre eles, caso exista. A essa combinação de características dá-se o nome de fusão multimodal. A fusão de diferentes modalidades pode ser classificada de acordo com o momento em que ela é realizada: fusão prévia e fusão tardia (ATREY *et al.*, 2010).

Na fusão prévia os vetores de características são combinados produzindo um único vetor de características que abrange a informação combinada desses vetores. O vetor fundido é usado como entrada na etapa de decisão, produzindo uma única resposta final. Esse esquema é ilustrado na Figura 5. Com essa abordagem, é possível detectar e evidencar a correlação intermodal, possibilitando assim uma melhoria na acurácia da tarefa. Por apenas uma etapa de decisão ser requerida, também ocorre melhoria no desempenho. Contudo, isso só é válido se a complexidade computacional do processo de fusão for menor ou igual do que a realização da etapa de decisão para cada um dos vetores de características.

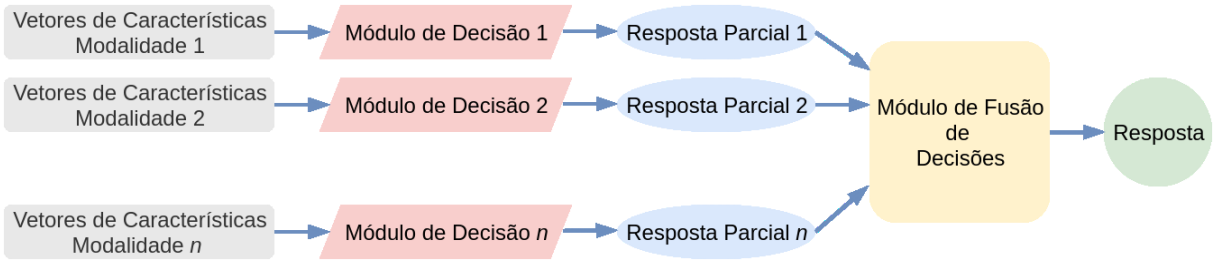
Figura 5 – Esquema de fusão prévia



Fonte: Kishi (2020).

Já na fusão tardia, a etapa de decisão é processada individualmente para cada vetor de características. As respostas parciais são então usadas como entrada em uma etapa de fusão de decisão, que as funde produzindo uma única resposta consensual. Essa abordagem é exibida na Figura 6. Atrey *et al.* (2010) também reportou haver trabalhos que realizam fusão híbrida, na qual uma parte dos vetores de características é fundida, resultando em menos respostas parciais, e essas respostas são também fundidas resultando na única final.

Figura 6 – Esquema de fusão tardia



Fonte: Kishi (2020).

2.4 Avaliação de Tarefas de Análise de Vídeo

Tarefas de análise multimídia são processos que manipulam, gerenciam, mineram, compreendem e visualizam diferentes tipos de dados de uma maneira efetiva para resolver um problema do mundo real (POUYANFAR *et al.*, 2018). Dentre elas, existem tarefas de análise de vídeo, tais como a Segmentação Temporal de Vídeo em Cenas (STVC) e a classificação de vídeo, que se destacam por dar suporte a outras tarefas de análise de vídeo e, desse modo, são de particular interesse para esta pesquisa. Na tarefa da STVC. o objetivo consiste em definir temporalmente as fronteiras de cenas de um vídeo.

A avaliação de tarefas de análise de vídeo é realizada com o uso de uma base de dados anotada, chamada de base confiável ou *groundtruth*, e uma ou mais medidas de eficácia para a devida comparação (HIEMSTRA; KRAAIJ, 2007). O *groundtruth* consiste em, pelo menos, um conjunto de vídeos voltados para uma tarefa específica, junto das respostas corretas para essa tarefa. Ele também pode vir acompanhado de dados complementares provenientes de etapas intermediárias da tarefa em questão, tais como características de baixo ou médio nível que podem ser extraídas dos vídeos e usadas em estágios intermediários da tarefa.

Como exemplos de bases confiáveis publicamente disponíveis, podem ser citados os trabalhos de Baraldi, Grana e Cucchiara (2015a) e Rotman, Porat e Ashour (2017). Baraldi, Grana e Cucchiara (2015a) desenvolveram a *BBC Dataset*, uma base de vídeos para a tarefa de STVC. Os vídeos dessa base compreendem a primeira temporada de *BBC Planet Earth*, uma série de documentários que retrata hábitats do planeta Terra. Além dos vídeos, a base

Tabela 1 – Propriedades importantes da *BBC Dataset*

Episódio	**Duração**	**Tomadas**	**Cenas**
From Pole to Pole	00:49:15	450	66
Mountains	00:48:04	395	53
Fresh Water	00:41:17	535	63
Caves	00:48:55	393	57
Deserts	00:48:59	469	55
Ice Worlds	00:49:17	425	62
Great Plains	00:49:03	473	71
Jungles	00:49:14	461	65
Shallow Seas	00:49:14	368	62
Seasonal Forests	00:49:19	529	65
Ocean Deep	00:49:14	418	53
Total	08:51:51	4916	672

Tabela 2 – Propriedades importantes da *OVSD Dataset*

Vídeo	**Duração**	**Tomadas**	**Cenas**
Big Buck Bunny	00:09:56	138	15
Cosmos Laundromat	00:12:10	107	7
Elephants Dream	00:10:53	137	9
Sintel	00:14:48	151	8
Tears of Steel	00:12:14	146	11
Valkaama	01:33:05	635	51
1000 Days	00:43:40	423	22
Boy Who Never Slept	01:09:46	353	36
CH7	01:26:28	1279	44
Fires Beneath Water	01:16:06	301	62
Honey	01:26:49	386	20
Jathia's Wager	00:21:01	159	15
La Chute D'une Plume	00:10:23	81	10
Lord Meia	00:37:05	274	27
Meridian	00:11:58	58	9
Oceania	00:54:19	271	31
Pentagon	00:50:33	364	31
Route 66	01:43:25	1086	55
Seven Dead Men	00:57:03	162	34
Sita Sings The Blues	01:21:31	965	52
Star Wreck	01:43:31	1291	55
Total	17:26:44	8767	604

também contém anotações a respeito dos quadros inicial e final de cada tomada e as tomadas inicial e final de cada cena de seus vídeos. Ao todo, a base possui 4916 tomadas e 672 cenas. Rotman, Porat e Ashour (2017) propuseram a *Open Video Scene Detection Dataset* (OVSD), também para a STVC. Apesar de seus 21 vídeos estarem enquadrados na categoria filme, eles apresentam conteúdo variado, como animações 2D e 3D, vídeos filmados com equipamento amador bem como profissional. Assim como a BBC, essa base contém as anotações das tomadas

e cenas, totalizando 8767 tomadas e 604 cenas. As tabelas 1 e 2 apresentam mais detalhes sobre a estrutura temporal hierárquica de cada um dos vídeos dessas bases.

A avaliação de tarefas de análise de vídeo pode ser objetiva, quando decisões de usuários não são envolvidas, ou subjetiva, quando é baseada na qualidade da experiência percebida pelo usuário. A tarefa de STVC, por exemplo, pode ser avaliada objetivamente ao comparar as tomadas preditas como transição de cenas com as corretas, indicadas pelo *groundtruth*. A anotação das tomadas corretas se torna subjetiva visto a subjetividade de uma cena já discutida anteriormente. Uma abordagem de minimizar isso seria combinar anotações de cenas geradas por vários espectadores e representá-las em uma única consensual, assim aumentando a probabilidade de um espectador aleatório concordar com a anotação real final. Com essa anotação e as tomadas preditas, pode-se avaliar o quão eficaz é uma técnica de segmentação de vídeo em cenas com um conjunto de medidas, tais como Precisão e Abrangência.

As medidas Precisão (P, do inglês, *Precision*) e Abrangência (R, do inglês, *Recall*) advêm da área de Recuperação de Informação (RIJSBERGEN, 1979) e já foram as mais usadas de acordo com Fabro e Böszörmenyi (2013) na tarefa de STVC. São realizadas três contagens: verdadeiros positivos (V_p), falsos positivos (F_p) e falsos negativos (F_n). V_p corresponde à quantidade de tomadas de transição de cenas corretamente preditas, ou seja, as que estão na anotação de fronteiras real. F_p corresponde à quantidade de tomadas de transição de cenas erroneamente preditas, ou seja, que não estão na anotação de fronteiras real. E F_n corresponde à quantidade de tomadas de transição de cenas erroneamente não preditas, ou seja, que estão na anotação de fronteiras real mas não na predita.

Precisão é definida como a razão entre a quantidade de fronteiras de cenas corretamente preditas e a quantidade de fronteiras preditas. Abrangência corresponde à razão entre a quantidade de fronteiras de cenas corretamente preditas e a quantidade de fronteiras reais. Assim, o cálculo feito para computar essas métricas é exibido nas Equações 2.1 e 2.2.

$$P = \frac{V_p}{V_p + F_p} \tag{2.1}$$

$$R = \frac{V_p}{V_p + F_n} \tag{2.2}$$

Maximizar a precisão significa minimizar falsos positivos, e maximizar a abrangência significa minimizar falsos negativos. Para simplificar em uma comparação que capture ambas propriedades, essas medidas são combinadas em uma única, medida-F ou F_1, que corresponde à média harmônica entre elas, calculada conforme a Equação 2.3.

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{2.3}$$

A utilização desse conjunto de medidas precisa ser analisada a depender da tarefa a ser avaliada, pois seu uso pode incorrer em problemas que impactam na interpretação da avaliação.

O principal deles é que falsos positivos e falsos negativos são rigidamente avaliados. No caso da STVC, por exemplo, uma tomada predita como transição distante 2 tomadas da verdadeira é contado como um falso positivo da mesma forma que outra distante 20 tomadas. Em uma tentativa de minimizar esse rigor, Hanjalic, Lagendijk e Biemond (1999) propuseram a inclusão de um critério de tolerância à posição de uma fronteira predita. Especificamente, uma transição predita dentro de 3 tomadas de distância da real ainda seria considerada um verdadeiro positivo. Apesar de flexibilizar a rigorosidade dessas medidas, a comparação entre diferentes trabalhos se torna laboriosa, pois autores diferentes podem empregar critérios diferentes. Além disso, Sidiropoulos *et al.* (2012) mostraram que essa F_1 não é simétrica em relação aos erros, ou seja, uma fronteira erroneamente predita a k tomadas além da correta pode ser avaliada diferentemente do que se fosse predita a k tomadas antes. Com essas motivações, pesquisadores sugeriram outras métricas de avaliação para a STVC, tais como o conjunto de medidas *Coverage* (C) e *Overflow* (O) (VENDRIG; WORRING, 2002; HAN; WU, 2011), e a Distância de Edição Diferencial (SIDIROPOULOS *et al.*, 2012).

Vendrig e Worring (2002) propuseram o conjunto de medidas *Coverage* (C) e *Overflow* (O) (Cobertura e Transbordamento, em português), também amplamente usado em avaliações de STVC. A *Coverage*, C_t, de uma cena real, $\tilde{s}_t$, mede a quantidade de tomadas dela agrupadas corretamente na predição, o que corresponde à maior sobreposição entre uma cena predita e $\tilde{s}_t$. Considerando $S = \{s_1, \ldots, s_{|S|}\}$ o conjunto de cenas preditas, $\tilde{S} = \{\tilde{s}_1, \ldots, \tilde{s}_{|\tilde{S}|}\}$ o conjunto de cenas reais e $|s_i|$ o número de tomadas da cena $|s_i|$, a Equação 2.4 exibe como a *Coverage* é calculada.

$$C_t = \frac{\max_{i=1}^{|S|} |s_i \cap \tilde{s}_t|}{|\tilde{s}_t|} \tag{2.4}$$

O *Overflow* O_t da cena $\tilde{s}_t$ mede quantas tomadas foram preditas além de suas fronteiras. A Equação 2.5 mostra como seu cálculo é realizado.

$$O_t = \frac{\sum_{i=1}^{|S|} |s_i \setminus \tilde{s}_t| \times \min(1, |s_i \cap \tilde{s}_t|)}{|\tilde{s}_{t-1}| + |\tilde{s}_{t+1}|} \tag{2.5}$$

Ao contrário das outras medidas, quanto menor o *Overflow*, melhor é a predição.

A *Coverage* e *Overflow* de um vídeo são calculadas com a média ponderada dessas medidas das cenas reais, conforme exibido na Equação 2.6, em que n é o número de tomadas do vídeo.

$$C = \sum_{t=1}^{|\tilde{S}|} C_t \times \frac{|\tilde{s}_t|}{n}, O = \sum_{t=1}^{|\tilde{S}|} O_t \times \frac{|\tilde{s}_t|}{n} \tag{2.6}$$

Assim como é feito no conjunto de medidas P e R, Baraldi, Grana e Cucchiara (2015b) sugeriram calcular a F_1 de *Coverage* e *Overflow* por meio da média harmônica de C e $1 - O$.

Deficiências nesse conjunto de medidas foram apontadas por Sidiropoulos *et al.* (2012) e Baraldi, Grana e Cucchiara (2015a). Assim como acontece com o conjunto de medidas P e R, as

medidas C e O não são simétricas. Além disso, o *Overflow* falha em avaliar casos extremos de subsegmentação, que é quando há muito menos cenas preditas do que reais. Com cenas vizinhas muito grandes, seus valores de *Overflow* facilmente ultrapassam 1, fazendo com que a F_1 assuma valores acima de 1 ou negativos. Han e Wu (2011) propuseram uma correção que contorna esse problema e que foi usada em outros trabalhos (KISHI; TROJAHN; GOULARTE, 2019; PEI *et al.*, 2021). Essa correção é descrita na Equação 2.7.

$$O_t = 1 - \frac{|\tilde{s}_t|}{\sum_{i=1}^{|S|} |\tilde{s}_t| \times \min(1, |s_i \cap \tilde{s}_t|)} \tag{2.7}$$

CAPÍTULO

3

TRABALHOS RELACIONADOS

Este capítulo trata dos principais trabalhos relacionados à fusão de informação multimodal. Relatos de técnicas de fusão de informação multimodal de propósito geral, ou seja, independente de tarefa ou base de dados, não foram encontrados. Dada essa ausência, técnicas de tarefas de análise de vídeo que acabam empregando fusão de informação multimodal como uma de suas etapas são abordadas neste capítulo.

As primeiras abordagens de fusão prévia multimodal usavam a concatenação de vetores de características de baixo nível (SNOEK; WORRING; SMEULDERS, 2005). Essas abordagens passaram a empregar operadores mais avançados, tais como soma e histograma conjunto (WANG *et al.*, 2011; WANG *et al.*, 2017). Uma lacuna foi deixada ao se trabalhar com vetores de características de diferentes descritores em um espaço heterogêneo, pois a semântica intermodal era deixada de lado. Trabalhos posteriores perseguiram essa lacuna ao focar em desenvolver métodos para extrair semântica a partir de dados brutos e expressá-la melhor em representações multimodais fundidas. Esses esforços levaram a métodos complexos que acoplaram as etapas de processamento para fusão de informação, extração de dados e o algoritmo das tarefas propriamente ditas.

Zhang *et al.* (2016) corroboraram que características de baixo nível podem não oferecer representatividade suficiente para tarefas de análise de vídeo devido à lacuna semântica. Eles combinaram a semântica presente entre objetos, cenas e ações humanas em um *framework* de fusão profunda para a tarefa de reconhecimento de eventos. Contudo, maximizar a correlação entre características de modalidades diferentes, assim como decorrelacionar características que pertencem a classes distintas na mesma modalidade não se torna efetivo em tarefas de classificação quando aplicadas em um espaço de fusão de características tão heterogêneo.

Güder e Çiçekli (2017) aplicaram *Principal Componente Analysis* (PCA) em cada descritor e realizou a fusão tardia das saídas das *Support Vector Machines* (SVMs) em um esquema de fusão híbrida por meio de regras de associação. Conforme a quantidade de dados

multimídia se torna disponível, mais processamento esse tipo de abordagem demanda. Apesar da melhor performance, a etapa de fusão de informação nesses estudos foi feita em um espaço heterogêneo, ou seja, vetores de características de diferentes modalidades e/ou dimensionalidades e/ou níveis semânticos. As correlações intramodais e intermodais não são modeladas de forma otimizada por algoritmos de aprendizado de máquina ao se lidar com um grande número de amostras (YANG; LANG; SONG, 2021).

Outros trabalhos tentaram melhorar a expressividade da fusão ao modelar diferentes maneiras de mapear a semântica contida nos dados entre modalidades enquanto simultaneamente se movia o espaço heterogêneo para um espaço comum multimodal usando representações BoF (VEMBU *et al.*, 2013; JHUO *et al.*, 2014; KISHI; TROJAHN; GOULARTE, 2019). Vembu *et al.* (2013) mapearam correlações entre características de médio nível ao se calcular a similaridade com medidas alternativas como TF-IDF baseado em NPMI (do Inglês, *Normalized Pointwise Mutual Information*). Após construir a matriz do grafo de similaridade, um algoritmo de agrupamento espectral de grafos foi utilizado para identificar grupos de palavras. Nessa etapa, o agrupamento depende das características extraídas e da medida de similaridade usada, ocasionando uma maior ou menor eficácia na tarefa.

A técnica proposta por Jhuo *et al.* (2014) foi baseada em detecção de correlação para fundir características de médio nível das modalidades aural e visual. Nessa técnica, um grafo bipartido é construído no qual os vértices correspondem às características e as arestas correspondem ao relacionamento entre uma característica visual e aural. O peso de cada aresta é definido como uma medida de correlação entre as características ligadas. As palavras bimodais foram determinadas nos grupos após o algoritmo de particionamento de grafo bipartido. Contudo, essa técnica é limitada a duas modalidades, o que força à concatenação (ou outro pré-processamento) de alguns vetores de características de médio nível antes de aplicá-la. Além disso, também falha em detectar correlação intramodal visto que as arestas sempre ligam características de modalidades diferentes.

Seguindo um caminho diferente, Kishi, Trojahn e Goularte (2019) projetaram outra maneira para mapear correlações entre características de médio nível. Nela todas as representações de médio nível são agrupadas de acordo com sua coocorrência ao longo das tomadas do vídeo. Cada grupo define uma palavra multimodal, que é representada por algum tipo de *pooling*, como soma ou máximo. A limitação presente no trabalho de Jhuo *et al.* (2014) foi ultrapassada ao se permitir a formação de grupos com características oriundas de uma única modalidade, o que pode acabar evidenciando alguma correlação intramodal. Por depender de um algoritmo de agrupamento, esse passo requer algum ajuste fino para estimar os melhores parâmetros, que podem variar de acordo com as características trabalhadas.

Trabalhos recentes tentaram desenvolver *frameworks* que utilizam redes neurais profundas (RAO *et al.*, 2020; TROJAHN; GOULARTE, 2021; PEI *et al.*, 2021). Tomando a tarefa de STVC como exemplo, Rao *et al.* (2020) integraram informações de diferentes estruturas hierár-

quicas, como clipe, segmento e filme, em um *framework* de segmentação em cenas, modelado como um de classificação. Consequentemente, métricas de classificação, como a *Average Precision* foi utilizada para avaliar a eficácia da tarefa. Tal métrica não é adequada para avaliar com acurácia tarefas de segmentação. O desenho de um *framework* de classificação acabou acoplando a escolha da métrica de avaliação, interferindo na utilização de métricas mais apropriadas e independentes de modelo.

Trojahn e Goularte (2021) projetaram um *framework* profundo, no qual incluía extração de características e segmentação, especificamente para a tarefa de STVC. Foram extraídas características VGGNet e CSIFT para a modalidade visual, MFCC para a aural e Word2Vec para a textual. O algoritmo de segmentação para classificar se uma cena seria considerada de transição ou não foi uma combinação de Redes Neurais Convolucionais (CNN, do Inglês) e Redes Neurais Recorrentes (RNN, do Inglês). As saídas das CNNs eram concatenadas e tomadas como entrada para uma unidade Long-Short Term Memory (LSTM) da RNN, onde a fusão profunda ocorria e a classificação da tomada era dada como saída. De maneira similar, Pei *et al.* (2021) modelaram um algoritmo de segmentação baseado numa rede de grafos convolucionais em um *framework* que incluía extração de quadros-chave e agrupamento de tomadas. Apesar de esses trabalhos atingirem resultados do estado da arte, a etapa de segmentação requer uma etapa de treinamento de acordo com as características extraídas. Mudanças na base de dados, processo de extração ou método de fusão irão requerer uma etapa de retreinamento.

Lacunas

Como pode ser visto a partir de uma análise dos trabalhos relacionados, fusão de informação enfrenta problemas relacionados a heterogeneidade dos dados e diferentes dimensionalidades. O fato de se modelar um problema associado a uma tarefa como um *framework* agregado torna a tarefa um monólito. A expressividade das representações finais pode ter melhorado, mas ao custo de se tornar a etapa de fusão inseparável do processamento da semântica multimodal. Isso acaba limitando possíveis melhorias em um passo isolado. Além disso, qualquer mudança no *pipeline*, como uma base de dados diferente, extração de características diferentes ou uma mudança no algoritmo da tarefa propriamente dita acaba requerendo uma etapa extra de retreinamento ou remodelagem a depender das mudanças realizadas.

CAPÍTULO

4

OPERADORES DE FUSÃO PRÉVIA APLICADOS AO MÉDIO NÍVEL SEMÂNTICO

Um Operador de Fusão Prévia (OFP) consiste em uma sequência de operações que são aplicadas a um conjunto de vetores de características, sejam elas profundas ou *handcrafted*, para obter uma representação única e consensual desse conjunto.

Os vetores de características no espaço de características de médio nível são histogramas, diferente do espaço de baixo nível. Cada posição nesse histograma resulta de algum processamento sobre um conjunto de características de baixo nível. Com isso, pode-se definir o conceito de espaço de fusão. Sejam letras minúsculas com números em subscrito os histogramas *Bag-of-Features* (*BoF*) de médio nível, como v_1 and v_2, por exemplo. Sejam letras maiúsculas o vetor de características resultante do OFP, como V, por exemplo. Cada elemento é representada pelo seu respectivo índice. Assim, por exemplo, v_{23} representa a contagem de ocorrências do terceiro padrão latente na representação do vetor v_2. O espaço de fusão é a matriz M_{nk}, em que cada linha está associada a um histograma *BoF*. Dessa forma, a Equação 4.1 generaliza como um operador op é aplicado sobre os operandos v_1 e v_2, tendo V como o histograma resultante dessa operação.

$$V = v_1 \; op \; v_2 \tag{4.1}$$

Este capítulo descreve os OFPs propostos: Soma, Concatenação e Máximo.

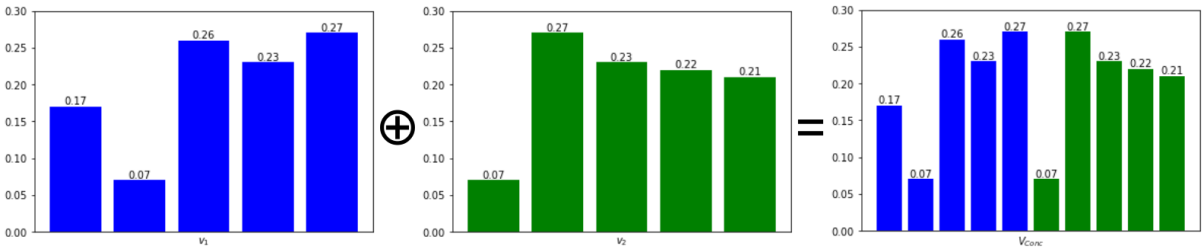
Concatenação

O operador Concatenação consiste em unir os histogramas lado a lado. Assim, a dimensão do vetor de características concatenado é a soma das dimensões dos seus operandos. Se cada um dos operandos for k-dimensional, então o vetor fundido será nk-dimensional. Ele será dado pela Equação 4.2.

$$V_{Conc} = v_1 \oplus v_2 \ldots \oplus v_n = [v_{11} \ldots v_{1k} \; v_{21} \ldots v_{2k} \ldots v_{n1} \ldots v_{nk}] \tag{4.2}$$

O operador de Concatenação mantém os valores originais de seus operandos ao custo de aumentar a dimensionalidade da representação. Isso pode facilitar a rastreabilidade para a informação original, o que contribui para análise da etapa de fusão. A Figura 7 demonstra essa operação. Nela o histograma resultante contém cada um dos seus dois operandos dispostos em sequência, um após o outro. Cada coluna ou posição refere-se a uma posição dos operandos, assim preservando a informação original, mas também dobrando a dimensionalidade.

Figura 7 – Exemplo de fusão pelo Operador Concatenação



Fonte: Elaborada pelo autor.

Soma

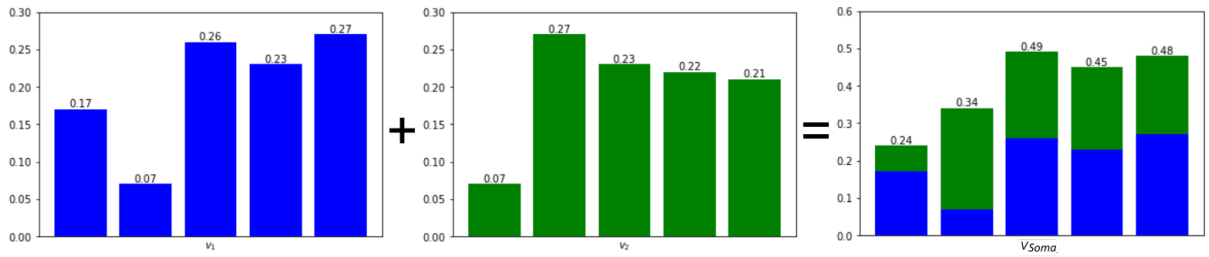
O operador Soma realiza a soma coluna a coluna dos histogramas *BoF*. Ele contém a restrição de que os operandos precisam ter o mesmo número de elementos, ou seja, a mesma dimensionalidade. No caso de características de médio nível, isso pode ser atingido se o número de grupos na etapa de agrupamento for a mesma para todas as modalidades, já que o tamanho do histograma *BoF* advém do número de grupos nessa etapa. Seu vetor de características fundido é representado pela Equação 4.3.

$$V_{Sum} = v_1 + v_2 \ldots + v_n = [\sum^{n} v_{i1} \sum^{n} v_{i2} \ldots \sum^{n} v_{ik}] \tag{4.3}$$

Como a soma é feita a cada coluna, a dimensão original dos operandos é mantida. Contudo, devido a maneira como o operador atua, quando ele é empregado em operandos de dimensões diferentes, alguma imputação é requerida. Geralmente a imputação costuma ser feita preenchendo-se zeros no fim dos operandos de menores dimensionalidades.

Em vez de aumentar a dimensionalidade como o operador de Concatenação, este operador aumenta a amplitude do sinal ao longo dos vetores de características. Altos valores de sinais existentes em um ou mais dos operandos podem favorecê-los na representação fundida mesmo quando há baixos valores de sinais nos outros operandos. Isso pode ajudar a destacar um sinal relevante presente em um sinal de baixo valor. Analogamente, o caso oposto representa uma desvantagem. Se há sinais irrelevantes com altos valores dentre um relevante, mas com baixo valor, que poderia ajudar a discernir melhor as características, esse sinal pode não ser

Figura 8 – Exemplo de fusão pelo Operador Soma



Fonte: Elaborada pelo autor.

adequadamente representado após a fusão com o operador Soma. A Figura 8 exemplifica essa operação. Cada coluna ou posição da representação resultante é a soma dos elementos nos operandos que estão nessa mesma coluna ou posição. Apesar da dimensionalidade ser preservada, a informação original, presente nos operandos, não pode mais ser localizada.

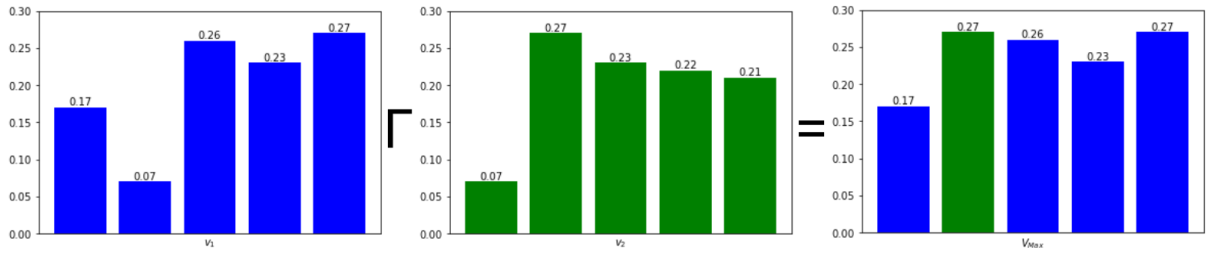
Máximo

O operador Máximo computa o valor máximo de cada coluna no espaço de fusão. Assim, seu vetor de características fundido é calculado conforme a Equação 4.4.

$$V_{Max} = v_1 \,\Gamma\, v_2 \,\ldots\, \Gamma\, v_n = [\max_{i=1}^{n} v_{i1} \; \max_{i=1}^{n} v_{i2} \,\ldots\, \max_{i=1}^{n} v_{ik}] \tag{4.4}$$

A Figura 9 ilustra um exemplo dessa operação. Cada coluna ou posição do vetor resultante representa o elemento de maior valor dentre os elementos dos operandos daquela mesma coluna ou posição. Ao invés de os sinais serem somados, como no operador Soma, um é preservado em detrimento de os outros serem descartados.

Figura 9 – Exemplo de fusão pelo Operador Máximo



Fonte: Elaborada pelo autor.

Apesar de não ser requerido um pré-processamento de imputação para usar esse operador, os operandos com maior tamanho determinarão a maioria dos elementos já que os com menores tamanhos não serão comparados nas últimas colunas, ou seja, as colunas em que eles não têm

representação. Além disso, por não haver de fato uma comparação nesses casos, a interpretação do resultado perde a semântica relativa à operação de máximo. Nesses cenários as modalidades com tamanhos maiores prevalecerão na representação resultante sobre as de tamanhos menores. Analogamente ao operador Soma, altos valores de sinais determinam o vetor resultante. Isso pode ajudar a representar melhor os valores dominantes nos operandos. Contudo, se alguns dos sinais dominantes forem irrelevantes, eles irão persistir ao longo da operação sendo levados para o resultado final. Isso pode afetar essa representação e, portanto, a eficácia da tarefa em questão.

CAPÍTULO

5

AVALIAÇÃO DA APLICAÇÃO DOS OFPS PARA STVC

Este capítulo descreve a análise comparativa realizada para medir a eficácia dos OFPs no espaço de características de nível médio semântico. A Seção 5.1 descreve a tarefa utilizada para essa análise, como as características foram extraídas e representadas, algoritmos utilizados e como a avaliação foi conduzida. Os resultados são apresentados e discutidos na Seção 5.2. E, por fim, uma análise comparativa da complexidade computacional dos OFPs e técnicas comparadas é apresentada e discutida na Seção 5.3.

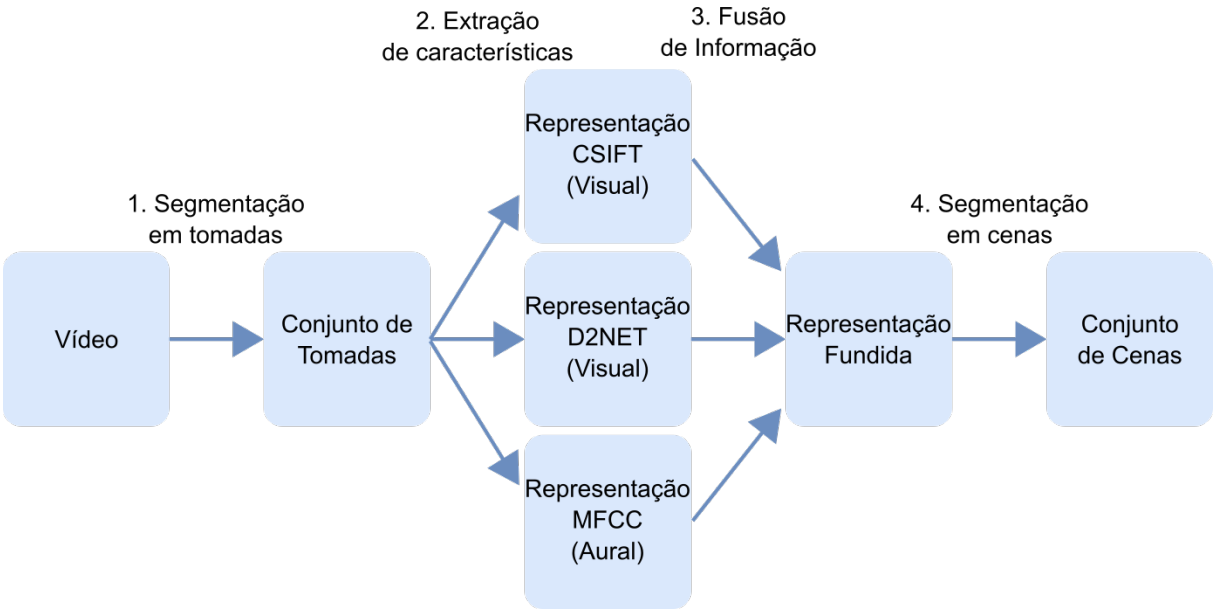
5.1 Metodologia

A segmentação temporal de vídeo em cenas foi utilizada como tarefa para medir a eficácia dos OFPs propostos. Essa tarefa é adequada para a análise proposta devido a um conjunto de razões. É uma tarefa conhecida, usada como pré-processamento para outras tarefas de análise de vídeo, como classificação e sumarização. É uma tarefa modular, permitindo que ela seja organizada em um *pipeline* de etapas de processamento, simplificando as mudanças e o acompanhamento onde possíveis ganhos de eficácia possam surgir. Também é possível encontrar trabalhos relacionados na literatura atual usando *datasets* e métricas de avaliação disponíveis e replicáveis, tornando possível a comparação com técnicas do estado da arte. As subseções seguintes (de 5.1.1 a 5.1.7) decrevem como as etapas do *pipeline* foram desenvolvidas.

5.1.1 Pipeline

A Figura 10 retrata o *pipeline* da STVC usada nos experimentos. Ele é composto por 4 etapas consecutivas, em que a saída de uma é usada como entrada para a seguinte. Este *pipeline* já foi aplicado no domínio da STVC e em trabalhos recentes com resultados do estado da arte

Figura 10 – *Pipeline* da segmentação temporal de vídeo em cenas utilizado



Fonte: Adaptada de Kishi (2020).

(BARALDI; GRANA; CUCCHIARA, 2015a; KISHI; TROJAHN; GOULARTE, 2019; PEI *et al.*, 2021).

A primeira etapa do *pipeline*, segmentação em tomadas, tem como objetivo detectar os quadros inicial e final de cada tomada do vídeo. Essa etapa não demandou processamento visto que os *datasets* utilizados já contêm anotações das fronteiras das tomadas. Essa abordagem é vantajosa pois isola a avaliação de eficácia de qualquer viés que possa ocorrer nessa etapa a depender do algoritmo de segmentação de tomadas aplicado.

5.1.2 Datasets e Ferramentas

As bases confiáveis *BBC Dataset* e *OVSD Dataset* foram utilizadas na avaliação da segmentação temporal de vídeo em cenas usando as representações fundidas geradas pelos operadores. Essas bases foram as escolhidas por serem recentemente utilizadas em trabalhos de STVC (RAO *et al.*, 2020; PEI *et al.*, 2021) e estarem disponíveis publicamente.

Os *scripts* para a execução do *pipeline* foram escritos na linguagem de programação *Python* por ela oferecer um conjunto de bibliotecas voltadas para a análise multimídia. A *OpenCV*[1] (*Open Source Computer Vision Library*) é uma biblioteca de processamento de imagem, que contém alguns extratores de características, como o SIFT, além de permitir o processamento de vídeo quadro a quadro. A *Python Speech Features*[2] contém uma implementação do MFCC, que foi usada nos experimentos. Uma implementação do CSIFT para *Windows* fornecida por Sande,

[1] <https://pypi.org/project/opencv-python/>
[2] <https://python-speech-features.readthedocs.io>

Gevers e Snoek (2011) foi utilizada. Como os OFPs propostos suportam vários operandos, uma característica profunda (*deep feature*) também foi empregada por poder aumentar a semântica contida na representação final. Para este cenário, a rede D2-Net foi utilizada usando seus parâmetros padrões em um modelo pré-treinado. Essa rede foi utilizada por extrair características locais, além de as representações obtidas por ela atingiram resultados competitivos nas tarefas de correspondência de imagens e reconstrução 3D (DUSMANU *et al.*, 2019).

5.1.3 Extração de Características

A extração de características é o segundo passo no *pipeline* adotado. Nessa etapa as características visuais e aurais dos vídeos foram extraídas. Na modalide visual um ou mais quadros-chave foram selecionados para representar cada tomada. O método de extração de quadros-chave utilizado foi o desenvolvido por Kishi, Trojahn e Goularte (2019). Assim, mantendo um mesmo processo de seleção de quadros-chave, um viés que possa ocorrer nessa etapa é eliminado. O quadro mais similar aos demais em uma tomada compõe o conjunto inicial de quadros-chave dessa tomada. Então, para adicionar variabilidade ao conjunto de quadros-chave, o quadro menos similar aos quadros já adicionados ao conjunto foi adicionado. Esse passo foi repetido até que não houvesse nenhum quadro na tomada que fosse menos similar aos quadros-chave do que um limiar, definido como 0,4 nos experimentos. Valores menores que 0,4 acabaram retornado um quadro-chave por tomada, ao passo que valores maiores que esse limiar retornaram uma alta quantidade de quadros-chave (KISHI, 2020). A medida de similaridade usada foi intersecção de histogramas de cor por sua comprovada eficácia em tarefas de comparação entre quadros-chave (HAN; WU, 2011; BARALDI; GRANA; CUCCHIARA, 2015b). O algoritmo de extração de quadros-chave é descrito no Algoritmo 2. A extração de características visuais é originada a partir dos quadros-chave para cada tomada.

Já a representação do conteúdo aural de uma tomada incluiu todo seu áudio já que os dados da modalidade aural são consideravelmente menores que os da visual. A extração de características aurais foi realizada pelo descritor MFCC em que o fluxo de áudio foi dividido em quadros de 30 milissegundos com janelas de sobreposição de 10 milissegundos. Essa abordagem provou ser satisfatória em aplicações de reconhecimento de fala (SEN; DUTTA; DEY, 2019). Além disso, MFCC é comumente utilizado em trabalhos do domínio de STVC (LOPES; TROJAHN; GOULARTE, 2014; ROTMAN; PORAT; ASHOUR, 2017; KISHI; TROJAHN; GOULARTE, 2019). Após a extração de características de baixo nível, elas foram processadas para se obter as características de médio nível.

5.1.4 Representação de Características em Médio Nível

A representação baseada em *BoF* foi aplicada para cada característica para extrair semântica latente contida nos vetores de característica de baixo nível e mover o espaço de representações heterogêneas unimodais para um comum. O processo para gerar os histogramas

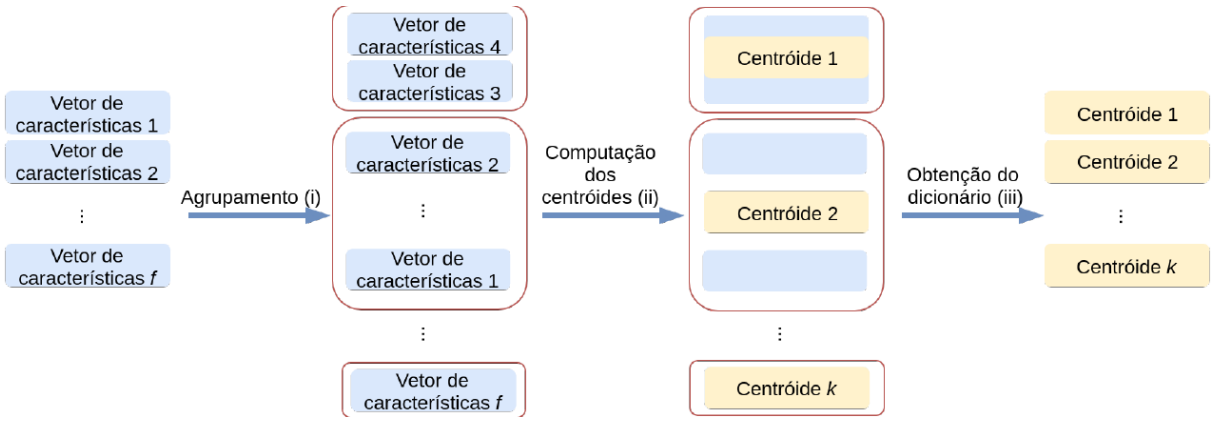
Algoritmo 2 – Extração de Quadros-chave

Entrada: Conjunto de quadros da tomada S; Limiar l; Função de similaridade σ
Retorno: Conjunto de quadros-chave K

$K \leftarrow \emptyset$
$K \leftarrow K \cup \{c\}$, sendo $c \in S$ o quadro cuja média de valores de similaridade a todos os outros quadros em S é máxima
$S \leftarrow S \setminus \{c\}$
$c \leftarrow d$, sendo $d \in S$ o quadro cuja média de valores de similaridade a todos os outros quadros em K é mínima
$avg\_sim \leftarrow \sum_{k \in K} \sigma(c,k)/|K|$
enquanto $|S| \neq 0$ & $avg\_sim \leq l$ **faça**
    $K \leftarrow K \cup \{c\}$
    $S \leftarrow S \setminus \{c\}$
    $c \leftarrow d$, sendo $d \in S$ o quadro cuja média de valores de similaridade a todos os outros quadros em K é mínima
    $avg\_sim \leftarrow \sum_{k \in K} \sigma(c,k)/|K|$
fim enquanto
retorna K

de frequência baseados em *BoF* está discutido em detalhes na Subseção 2.2.2 e ilustrado nas Figuras 11 e 12.

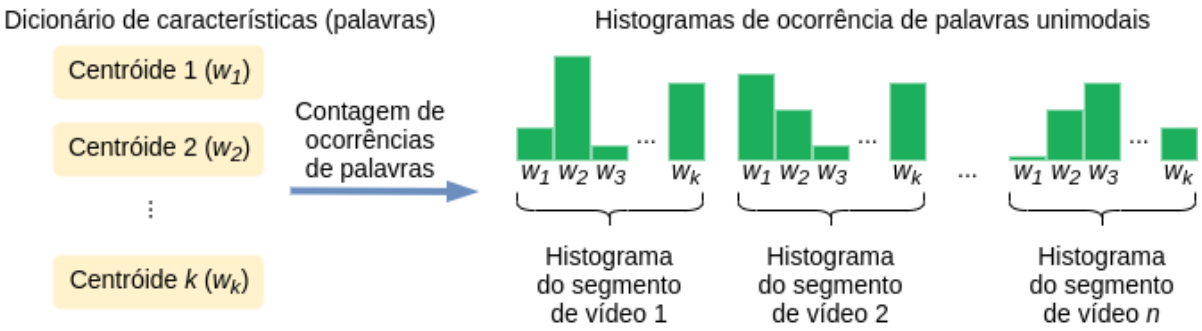
Figura 11 – Obtenção da *Bag-of-Features*



Fonte: Kishi (2020).

O agrupamento dos vetores de característica foi realizado pelo algoritmo *k-means++*, presente na biblioteca *Scikit-learn* (PEDREGOSA *et al.*, 2011) do *Python*. O número de clusters adotado foi 100, mantendo os demais parâmetros do algoritmo padronizados pela biblioteca, assim replicando a configuração utilizada por Kishi (2020), por já ser uma configuração empiricamente testada em trabalhos anteriores (KISHI; TROJAHN; GOULARTE, 2018; KISHI; TROJAHN; GOULARTE, 2019; KISHI, 2020). Com os clusters gerados, os centróides foram utilizados para compor os dicionários de cada característica, conforme o processo descrito na Figura 12.

Figura 12 – Obtenção do histograma de frequências *Bag-of-Features*



Fonte: Kishi (2020).

Apesar de histogramas *BoF* de diferentes modalidades representarem a contagem de ocorrência de padrões e a mesma dimensão, quando gerados com o mesmo número de clusters, os padrões latentes contidos em cada coluna são distintos e não correlacionados. Uma comparação direta entre histogramas *BoF* de diversas modalidades ou descritores não possui semântica embutida. Devido a isso, uma etapa de fusão prévia deve ser aplicada para transferir o espaço heterogêneo de informações unimodais para um comum multimodal.

5.1.5 Fusão de Informação

Os OFPs propostos foram aplicados na etapa de fusão prévia, ou seja, após os vetores de características de médio nível, os histogramas *BoF* aurais e visuais, serem gerados. Com esses operadores aplicados nesses histogramas e normalizando os resultados pela norma l_1, o *pipeline* segue para a próxima etapa, o algoritmo de segmentação em cenas.

5.1.6 Segmentação em Cenas

Os histogramas fundidos foram usados como entrada para o algoritmo de segmentação em cenas. O algoritmo STG (*Scene Transition Graph*) (YEUNG; YEO; LIU, 1998) foi utilizado por seu uso comum em trabalhos do domínio de STVC, como em Baraldi, Grana e Cucchiara (2015a) e Kishi, Trojahn e Goularte (2019), permitindo uma melhor comparação entre os operadores e técnicas.

Nesse algoritmo as tomadas são agrupadas utilizando o algoritmo de agrupamento hierárquico *complete linkage* que foi adaptado para a tarefa de segmentação para evitar a inclusão de tomadas temporalmente distantes caso elas fossem similares (YEUNG; YEO; LIU, 1998). Isso foi garantido pelo cálculo da distância entre tomadas, $\hat{d}_{max}$, dado pela Equação 5.1, em que C_i e C_j são grupos de tomadas e $\hat{d}(S_i, S_j)$ é dado pela Equação 5.2.

$$\hat{d}_{max}(C_i, C_j) = \max_{x \in C_i, y \in C_j} \hat{d}(x, y) \tag{5.1}$$

$$\hat{d}(S_i,S_j) = \begin{cases} d(S_i,S_j), & \text{se } d_t(S_i,S_j) \leq T \\ \infty, & \text{caso contrário} \end{cases} \tag{5.2}$$

A medida de distância adotada no algoritmo é representada por d, d_t é a distância temporal entre tomadas e w é uma janela temporal. O critério de parada é determinado pelo parâmetro δ. Quando $\hat{d}_{max}(A,B) > \delta$, $\forall$ pares de grupos (A,B) com $A \neq B$, o agrupamento termina. Os grupos obtidos correspondem aos vértices do grafo de transição de cenas. Uma aresta entre os vértices A e B é criada se e somente se o grupo A contém uma tomada temporalmente adjacente a qualquer tomada do grupo B.

Algoritmo 3 – Segmentação em Cenas STG

Entrada: Conjunto $S = \{s_1,...,s_n\}$ de tomadas do vídeo; Conjunto de vetores de características $H = \{h_1,...,h_n\}$ referentes a cada uma das tomadas em S; Função de distância d; Janela temporal w; Limiar de agrupamento δ

Retorno: Conjunto $B = \{b_1,...,b_k\}$ de pares de índices das tomadas inicial e final de cada cena

1: Efetue um agrupamento *complete linkage* sobre o conjunto S por meio de suas representações em H usando a distância d. Considere $d(h_i,h_j) = \infty$ se $|s_j - s_i| > w$. Encerre quando todas as distâncias entre os grupos forem maiores do que δ. Atribua os grupos resultantes ao conjunto C
2: Crie um grafo $G = \{V,E\}$
3: $V \leftarrow$ índices de C
4: Insira arestas $e = (a,b)$ em E, com a,$b \in V$ para todos os pares de tomadas $s_i \in c_a$ e $s_{i+1} \in c_b$ com $\{s_i,s_{i+1}\} \subset S$ e $\{c_a,c_b\} \subset C$
5: Remova as arestas de corte em G e atribua os subgrafos resultantes ao conjunto R
6: $B \leftarrow \emptyset$
7: **para** cada grafo $r \in R$ **faça**
8: $\quad r_f \leftarrow min(r)$
9: $\quad r_l \leftarrow max(r)$
10: $\quad B \leftarrow B \cup \{\{r_f,r_l\}\}$
11: **fim para**
12: Ordene B
13: **retorna** B

As arestas de corte no grafo determinam transições de cena. Cada subgrafo consiste em uma cena. Tomadas no mesmo grupo possuem interações confinadas a esse grupo exceto com uma transição de cenas. Arestas de corte têm alta probabilidade de corresponderem a transições entre grupos, ou seja, as cenas referentes a esses grupos. O Algoritmo 3 descreve a segmentação em cenas feita pelo STG. Uma medida de distância comumente adotada em segmentação de vídeo em cenas é a distância cosseno (YEUNG; YEO; LIU, 1998; BARALDI; GRANA; CUCCHIARA, 2015a; KISHI; TROJAHN; GOULARTE, 2019).

Os parâmetros w e δ do algoritmo STG para a base *BBC* foram definidos como 7 e 0,35, respectivamente. Esses valores foram empiricamente demonstrados como os melhores para essa base (KISHI, 2020). Na *OVSD*, devido à sua diversidade e à particularidade de cada vídeo, os

Tabela 3 – Melhores parâmetros do STG na *OVSD Dataset*

#	Vídeo	δ	w
01	1000 Days	0,45	25
02	Big Buck Bunny	0,45	15
03	Boy Who Never Slept	0,25	25
04	CH7	0,75	20
05	Cosmos Laundromat	0,40	20
06	Elephants Dream	0,45	15
07	Fires Beneath Water	0,40	10
08	Honey	0,75	20
09	Jathia's Wager	0,80	12
10	La Chute D'une Plume	0,80	25
11	Lord Meia	0,40	25
12	Meridian	0,65	15
13	Oceania	0,55	20
14	Pentagon	0,25	25
15	Route 66	0,35	35
16	Seven Dead Men	0,10	10
17	Sintel	0,80	25
18	Sita Sings The Blues	0,25	35
19	Star Wreck	0,40	20
20	Tears of Steel	0,50	20
21	Valkaama	0,55	25

parâmetros foram determinados empiricamente visando aos melhores resultados nas técnicas do estado da arte, como a M4InFus (KISHI, 2020). A Tabela 3 lista a configuração de parâmetros para cada vídeo.

Com as segmentações dos vídeos realizadas e obtidas as fronteiras das cenas, a eficácia da tarefa pode ser avaliada por um conjunto de medidas, conforme descrito na Subseção 5.1.7.

5.1.7 Avaliação

A eficácia dos OFPs propostos na tarefa de STVC foi avaliada usando o conjunto de medidas *Coverage* (C), *Overflow* (O) e sua F_1. Apesar de as medidas Precisão e Abrangência terem sido utilizadas em trabalhos anteriores de STVC, elas foram originárias da área de Recuperação da Informação, voltadas à análise de tarefas de classificação. Assim, acabam não sendo as mais apropriadas para a análise dos resultados da segmentação. Por isso, C e O foram propostas visando a contornar esse problema de imprecisão na análise (VENDRIG; WORRING, 2002). Esse conjunto de medidas avalia, baseado na quantidade de tomadas, o quanto uma cena predita pelo algoritmo se sobrepõe com a cena real. A equação adaptada por Han e Wu (2011) foi utilizada neste trabalho para evitar valores negativos de T nos casos de subsegmentação que pudessem levar a análises equivocadas. Como baixos valores de O são melhores, a média harmônica, ou seja, a F_1 desse conjunto de medidas, foi calculada entre os valores de C e $1 - O$

para cada vídeo. Essas medidas foram calculadas por um *script* Python que tem como entrada um arquivo no formato csv contendo as anotações de início e fim das cenas de um vídeo, e a segmentação obtida pelo algoritmo. Isso foi feito para cada vídeo de cada base e para cada operador.

Além da avaliação da eficácia da tarefa, os resultados entre as diferentes técnicas do estado da arte e os operadores foram avaliados quanto à sua diferença ou similaridade com uso do teste *t de Student*. Esse teste verifica se as médias de conjuntos diferentes são de fato diferentes com algum nível de significância estatística. Para isso, a normalidade dos dados e a igualdade de variâncias são assumidas. A normalidade pode ser testada pelo teste de SHAPIRO e WILK (1965). Já a igualdade de variâncias, pelo teste de Levene (1961). Caso a igualdade de variâncias não possa ser assumida, uma versão do *t de Student* para dados com variâncias diferentes, o teste de WELCH (1947) pode ser utilizada. As implementações desses testes podem ser encontradas na biblioteca *SciPy* para *Python* (VIRTANEN *et al.*, 2020). Um p-valor retornado pelo teste *t de Student* ou de Welch acima de $\alpha\%$ indica que as médias das duas amostras são distintas com $1 - \alpha\%$ de significância estatística.

5.2 Resultados

5.2.1 BBC Dataset

A Tabela 4 apresenta os valores de C, O e F_{CO} obtidos para cada vídeo da *BBC Dataset* usando as representações geradas pelos operadores. Os baixos valores de *Overflow* indicam bons resultados para essa medida quando avaliada individualmente.

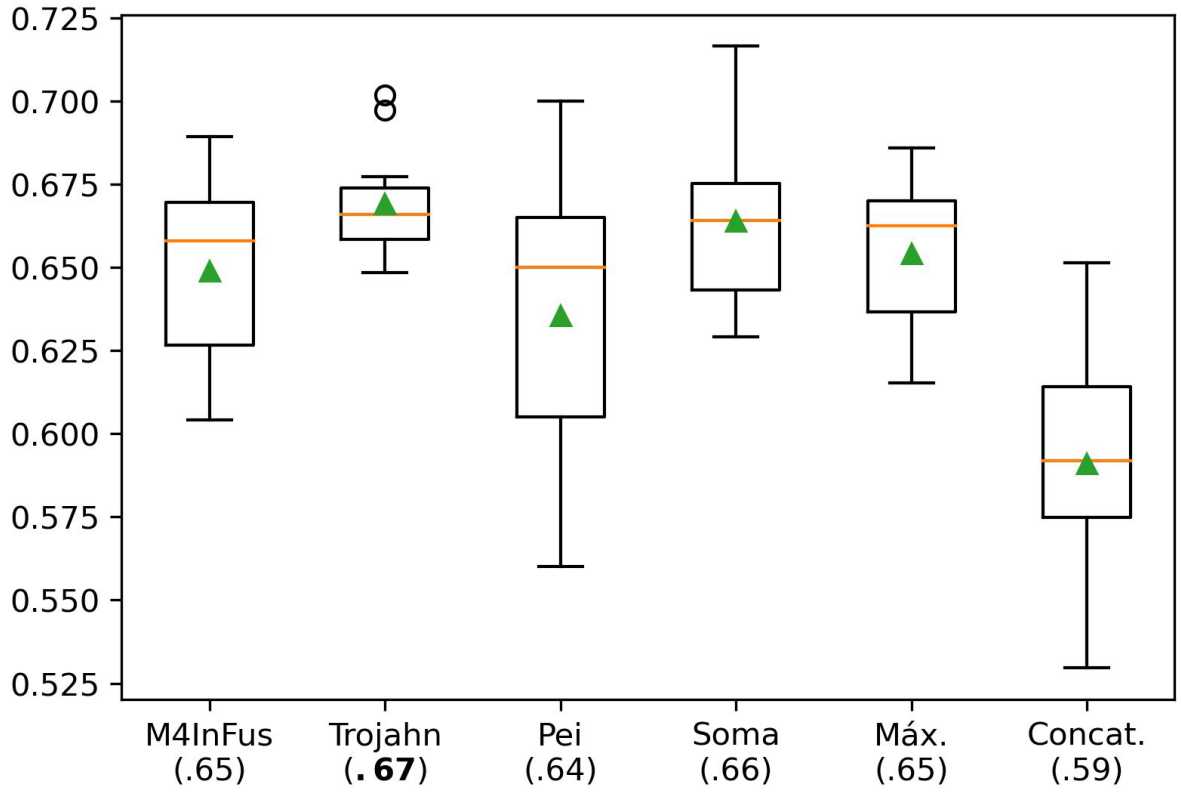
Tabela 4 – Valores de C, O e F_{CO} obtidos com os OFPs na *BBC Dataset*

Video	Soma			Concatenação			Máximo		
	C (%)	O (%)	F_{CO} (%)	C (%)	O (%)	F_{CO} (%)	C (%)	O (%)	F_{CO} (%)
01	55,11	26,71	62,91	45,33	15,70	58,96	55,56	27,23	63,01
02	53,42	18,09	64,67	46,58	11,08	61,14	53,92	22,08	63,74
03	59,29	18,04	68,81	40,00	3,57	56,54	58,35	21,40	66,98
04	61,31	23,59	68,03	49,47	13,15	63,04	59,62	19,28	68,58
05	58,13	22,58	66,40	54,01	17,99	65,13	57,70	21,55	66,49
06	60,87	32,57	63,98	46,88	19,72	59,19	56,14	26,71	63,58
07	57,41	19,84	66,91	36,59	4,16	52,96	52,16	9,25	66,25
08	55,02	16,44	66,35	42,11	4,55	58,43	50,96	12,53	64,40
09	68,66	25,09	71,65	46,59	8,77	61,68	61,58	24,28	67,92
10	53,94	22,62	63,57	37,66	7,88	53,46	49,11	17,64	61,53
11	55,65	15,78	67,02	43,92	7,15	59,64	53,52	10,34	67,03
Média	58,08	21,94	66,39	44,47	10,34	59,11	55,33	19,30	65,41
σ	4,21	4,79	2,44	4,92	5,39	3,56	3,64	5,95	2,16

Tabela 5 – Valores de F_{CO} obtidos na *BBC Dataset*

Video	M4InFus (%)	Trojahn (%)	Pei (%)	Soma (%)	Concatenação (%)	Máximo (%)
01	62	66	**67**	63	59	63
02	64	65	**68**	65	61	64
03	66	66	60	**69**	57	67
04	68	**70**	65	68	63	69
05	**68**	66	56	66	65	66
06	60	**65**	**65**	64	59	64
07	66	**68**	66	67	53	66
08	63	67	**70**	66	58	64
09	69	67	61	**72**	62	68
10	61	**70**	64	64	53	62
11	66	**67**	57	**67**	60	**67**
Média	65	**67**	64	66	59	65
σ	2,84	1,71	0,04	2,44	3,56	2,16

Figura 13 – Boxplot da F_{CO} na *BBC Dataset*



Fonte: Elaborada pelo autor.

A Tabela 5 reproduz os valores de F_{CO} obtidos com os OFPs e técnicas do estado da arte (KISHI; TROJAHN; GOULARTE, 2019; PEI *et al.*, 2021; TROJAHN; GOULARTE, 2021) na *BBC Dataset*. As técnicas do estado da arte superaram os operadores em 8 dos 11 vídeos dessa base. Contudo, os operadores Soma e Máximo alcançaram resultados próximos à técnica

de Trojahn e Goularte (2021), a que se saiu melhor na base, com menos de 2% de diferença. Os resultados de F_{CO} obtidos foram plotados em um *boxplot* na Figura 13. Nesse tipo de gráfico, os triângulos verdes representam a média de um conjunto de dados; a linha contínua em laranja, a respectiva mediana; e *outliers* são representados como círculos não preenchidos. Nota-se que os intervalos interquartis dos operadores Soma e Máximo são menores que todos, exceto a técnica de Trojahn e Goularte (2021). Isso mostra que esses operadores apresentaram menor variância nos resultados do que os demais.

Tabela 6 – Teste de Shapiro-Wilk na *BBC Dataset*

Técnica/Operador	**p-valor**
M4InFus	,481000
Pei	,605000
Trojahn	,147000
Soma	,707000
Máximo	,598000
Concatenação	,802000

Tabela 7 – Teste de Levene na *BBC Dataset*

Par de amostras	**p-valor**
M4InFus x Soma	,481000
M4InFus x Máximo	,443000
M4InFus x Concatenação	,724000
M4InFus x Pei	,353000
M4InFus x Trojahn	,077000
Soma x Máximo	,980000
Soma x Concatenação	,338000
Soma x Pei	,154000
Soma x Trojahn	,263000
Máximo x Concatenação	,311000
Máximo x Pei	,140000
Máximo x Trojahn	,236000
Concatenação x Pei	,555000
Concatenação x Trojahn	,068000
Pei x Trojahn	**,035000**

A Tabela 6 exibe os p-valores do teste de Shapiro-Wilk para os dados da *BBC Dataset* para cada operador ou técnica. Como todos apresentaram valores acima de 0,05, pode-se considerar que passaram nesse teste e que as amostras possuem distribuição normal. A Tabela 7 apresenta os p-valores do teste de Levene para os dados da *BBC Dataset* para cada par de operador ou técnica. Quase todos os pares manifestaram valores acima de 0,05, com exceção do par de técnicas de Pei *et al.* (2021) e Trojahn e Goularte (2021). Nesse par em particular, a versão adaptada do teste *t de Student*, o teste de Welch, foi aplicada.

A Tabela 8 expõe os p-valores obtidos. Os pares de amostras que obtiveram valores abaixo de 0,01 estão destacados em negrito. Esse resultado aponta que as médias para essas

amostras foram estatisticamente diferentes com 99% de nível de significância. Os demais pares de amostras apresentaram p-valores acima de 0,01. Assim, pode-se inferir que esses pares de médias foram estatisticamente similares com 99% de nível de significância. Nota-se que o resultado para o operador de Concatenação foi o pior dentre os comparados e sua média foi significativamente diferente das demais. Além disso, foi o que apresentou maior variância ao longo dos vídeos da base. Isso se comprova por ter o maior desvio padrão apresentado. Dentre os operadores, Soma e Máximo foram os que apresentaram melhores resultados e obtiveram médias significativamente similares ao melhor resultado, a técnica de Trojahn.

Tabela 8 – Teste de *t de Student*/Welch na *BBC Dataset*

Par de amostras	**p-valor**
M4InFus x Soma	,215000
M4InFus x Máximo	,641000
M4InFus x Concatenação	**,001000**
M4InFus x Pei	,424000
M4InFus x Trojahn	,065000
Soma x Máximo	,352000
Soma x Concatenação	**,000003**
Soma x Pei	,083000
Soma x Trojahn	,594000
Máximo x Concatenação	**,001000**
Máximo x Pei	,234000
Máximo x Trojahn	,098000
Concatenação x Pei	,020000
Concatenação x Trojahn	**,000004**
Pei x Trojahn	,038000

5.2.2 OVSD Dataset

A Tabela 9, por sua vez, apresenta os valores obtidos para a *OVSD Dataset*. Uma exceção peculiar para as sobresegmentações foi o resultado do operador Máximo para o vídeo 10 da *OVSD Dataset*. O vídeo possui 10 cenas e o algoritmo de segmentação detectou apenas 3. Nos casos de subsegmentação, altos valores de *Overflow* podem vir acompanhados também de altos valores de *Coverage*. Isso pode apontar que, apesar de poucas cenas terem sido detectadas, a maioria dessas poucas cenas tiveram suas fronteiras corretamente apontadas.

Tabela 9 – Valores de C, O e F_{CO} obtidos com os OFPs na *OVSD Dataset*

Video	Soma			Concatenação			Máximo		
	C (%)	O (%)	F_{CO} (%)	C (%)	O (%)	F_{CO} (%)	C (%)	O (%)	F_{CO} (%)
01	71,09	46,39	61,13	68,96	35,66	66,57	67,30	35,28	65,98
02	71,01	42,99	63,25	56,52	27,72	63,44	68,84	35,15	66,78
03	65,06	32,45	66,28	53,41	19,51	64,21	65,06	34,88	65,09
04	74,08	23,56	75,24	72,83	27,77	72,53	78,86	33,57	72,11
05	65,42	9,82	75,83	33,64	5,75	49,59	66,36	12,76	75,38
06	69,85	42,06	63,34	58,09	21,49	66,77	69,12	38,33	65,18
07	79,67	45,92	64,42	73,33	36,62	68,00	77,67	39,72	67,88
08	73,18	34,11	69,34	69,27	30,84	69,22	72,14	27,57	72,28
09	70,89	28,49	71,20	66,46	21,45	72,00	66,46	27,46	69,37
10	85,00	49,02	63,74	70,00	14,75	76,87	100,00	47,70	68,68
11	76,19	37,23	68,83	79,49	29,67	74,63	84,98	38,70	71,22
12	81,36	32,04	74,06	77,97	24,46	76,73	81,36	32,04	74,06
13	64,58	28,11	68,04	64,21	29,19	67,35	66,79	32,70	67,05
14	74,10	32,70	70,54	66,12	23,39	70,98	76,03	31,89	71,85
15	67,25	30,96	68,13	62,73	21,23	69,84	67,80	35,49	66,12
16	68,32	26,04	71,03	64,60	24,34	69,69	67,70	29,65	69,00
17	92,72	38,55	73,91	70,20	7,61	79,78	86,09	39,76	70,88
18	74,17	17,49	78,12	59,75	10,09	71,79	67,74	14,49	75,59
19	59,33	31,40	63,63	45,70	14,07	59,67	62,43	30,94	65,58
20	85,81	30,92	76,54	60,14	14,05	70,76	70,27	22,68	73,63
21	81,16	34,87	72,27	76,14	28,00	74,01	77,71	32,78	72,09
Média	73,82	33,10	69,47	64,26	22,27	69,26	73,37	32,07	69,80
σ	7,97	9,28	4,91	10,64	8,46	6,41	8,88	7,89	3,33

A Tabela 10 exibe os valores de F_{CO} obtidos com os OFPs e técnicas do estado da arte (KISHI; TROJAHN; GOULARTE, 2019; PEI *et al.*, 2021; TROJAHN; GOULARTE, 2021) na *OVSD Dataset*. As técnicas do estado da arte superaram os operadores em 18 dos 21 vídeos dessa base. Mesmo assim, o operador Máximo alcançou resultados superiores a uma das técnicas, a de Pei. A Figura 14 exibe os resultados de F_{CO} da *OVSD Dataset* em um *boxplot*. As médias dos resultados dos operadores foram próximas à média da melhor técnica e superaram as médias das outras duas. Os intervalos interquartis dos operadores Máximo e Concatenação foram menores do que os outros métodos, o que mostra que esses operadores apresentaram um comportamento mais estável nos resultados do que os obtidos pelas técnicas do estado da arte.

Tabela 10 – Valores de F_{CO} obtidos na *OVSD Dataset*

Video	M4InFus (%)	Trojahn (%)	Pei (%)	Soma (%)	Concatenação (%)	Máximo (%)
01	70	50	**71**	62	67	66
02	**67**	60	65	63	63	**67**
03	**72**	54	70	66	64	65
04	**76**	59	68	75	73	72
05	76	**83**	78	76	50	75
06	**69**	32	65	63	67	65
07	**70**	64	58	64	68	68
08	**75**	64	73	69	69	72
09	**76**	**76**	72	71	72	69
10	**80**	73	71	64	77	69
11	**75**	56	67	69	**75**	71
12	74	66	**86**	74	77	74
13	**68**	66	65	**68**	67	67
14	73	**75**	60	71	71	72
15	**72**	67	64	68	70	66
16	71	67	**74**	71	70	69
17	**88**	68	64	74	80	71
18	**79**	51	67	78	72	76
19	**69**	64	62	64	60	66
20	68	71	60	**77**	71	74
21	74	59	**79**	72	74	72
Média	**73**	63	69	69	69	70
σ	4,81	10,81	0,07	4,91	6,41	3,33

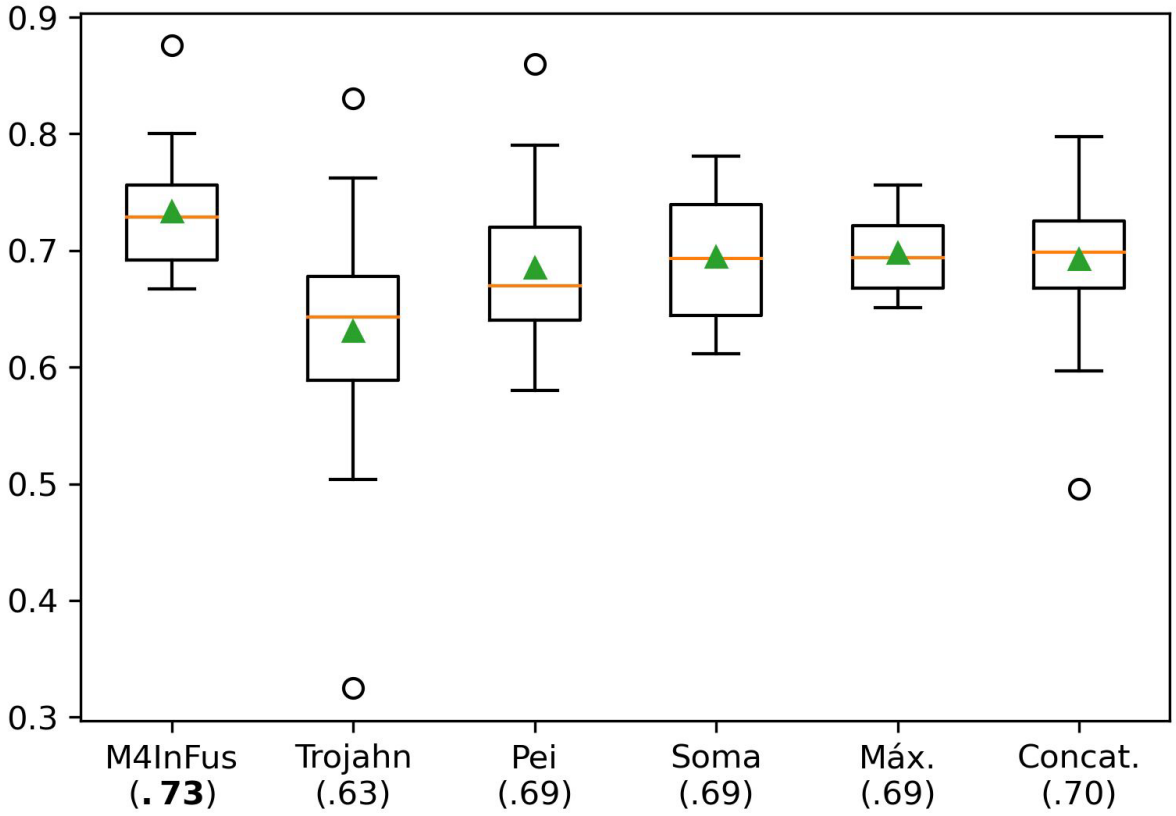
Tabela 11 – Teste de Shapiro-Wilk na *OVSD Dataset*

Técnica/Operador	**p-valor**
M4InFus	,055000
Pei	,437000
Trojahn	,329000
Soma	,445000
Máximo	,183000
Concatenação	,088000

A Tabela 11 reproduz os p-valores do teste de Shapiro-Wilk para os dados da *OVSD Dataset* para cada operador ou técnica. Como todos apresentaram valores acima de 0,05, pode-se considerar que passaram nesse teste e que as amostras possuem distribuição normal. A Tabela 12 ilustra os p-valores do teste de Levene para os dados da *OVSD Dataset* para cada par de operador ou técnica. Quase todos os pares refletiram valores acima de 0,05, com exceção dos pares M4Infus e Trojahn, Máximo e Pei, e Máximo e Trojahn. Nesses pares em específico, a versão adaptada do teste *t de Student*, o teste de Welch, foi aplicada.

A Tabela 13 expõe os p-valores obtidos. Os pares de amostras que obtiveram valores abaixo de 0,01 estão destacados em negrito. Esse resultado aponta que as médias para essas

Figura 14 – Boxplot da F_{CO} na *OVSD Dataset*



Fonte: Elaborada pelo autor.

amostras foram estatisticamente diferentes com 99% de nível de significância. Os demais pares de amostras apresentaram p-valores acima de 0,01. Apesar de a melhor técnica nessa base, a M4InFus, ter superado os operadores com diferenças de até 4%, seus resultados se mostraram estatisticamente similares aos dos operadores Soma e Concatenação com 99% de nível de significância. Os resultados que apresentaram diferenças significativas foram entre a técnica M4InFus e a técnica de Trojahn e o operador Máximo.

Todos os 3 operadores apontaram resultados estatisticamente similares entre eles. Em vídeos de assuntos mais diversificados e diferentes categorias, os OFPs atingiram resultados próximos entre si, com uma diferença de até 1%. Isso pode implicar que a escolha de um OFP na etapa de fusão pode acabar não sendo uma decisão relevante em bases maiores e mais variadas. Quando em um domínio específico, como documentários, por exemplo, as representações obtidas pelos operadores e seus resultados podem diferir mais entre si. Comparado à *BBC Dataset*, o operador Concatenação obteve melhores resultados na *OVSD Dataset*, que é uma base que contém vídeos mais diversos e com diferentes tipos de conteúdos. Isso pode suscitar a hipótese de que a operação de Concatenação pode ter resultados limitados em domínios específicos, como o de documentários, que é o domínio da *BBC Dataset*.

Tabela 12 – Teste de Levene na *OVSD Dataset*

Par de amostras	**p-valor**
M4InFus x Soma	,632000
M4InFus x Máximo	,295000
M4InFus x Concatenação	,520000
M4InFus x Pei	,198000
M4InFus x Trojahn	**,033000**
Soma x Máximo	,074000
Soma x Concatenação	,757000
Soma x Pei	,324000
Soma x Trojahn	,051000
Máximo x Concatenação	,144000
Máximo x **Pei**	**,029000**
Máximo x Trojahn	**,007000**
Concatenação x Pei	,586000
Concatenação x Trojahn	,113000
Pei x Trojahn	,228000

Tabela 13 – Teste de *t de Student*/Welch na *OVSD Dataset*

Par de amostras	**p-valor**
M4InFus x Soma	,015000
M4InFus x Máximo	**,009000**
M4InFus x Concatenação	,027000
M4InFus x Pei	,013000
M4InFus x Trojahn	**,001000**
Soma x Máximo	,805000
Soma x Concatenação	,908000
Soma x Pei	,617000
Soma x Trojahn	,019000
Máximo x Concatenação	,739000
Máximo x Pei	,454000
Máximo x Trojahn	,012000
Concatenação x Pei	,727000
Concatenação x Trojahn	,031000
Pei x Trojahn	,060000

5.3 Análise da Complexidade Computacional

Um OFP é mais eficiente computacionalmente comparado a uma fusão tardia quando a complexidade da fusão prévia e da tarefa aplicada à representação fundida é menor que a complexidade da tarefa aplicada a n vetores de características junto do algoritmo de fusão tardia adotado.

Seja $O_t(d)$ a complexidade computacional de uma tarefa de análise de vídeo sobre uma entrada de dimensionalidade d, e O_l a complexidade computacional para a etapa de fusão tardia, ou seja, o algoritmo utilizado para a fusão tardia. Com cada um dos n vetores de características

de um vídeo sendo k-dimensional, um OFP é mais eficiente computacionalmente comparado à fusão tardia quando a desigualdade 5.3 é obedecida.

$$O_e + O_t(d_e) < nO_t(k) + O_l \tag{5.3}$$

Na desigualdade 5.3, O_e representa a complexidade computacional desse OFP e d_e, a dimensionalidade do vetor de características resultante da aplicação do OFP. O primeiro membro dessa desigualdade remete à complexidade computacional associada ao OFP.

Com isso, se técnicas que empregam fusão prévia e/ou operadores produzem como resultado um vetor de características fundido com a mesma dimensionalidade, o termo $O_t(d_e)$ pode ser desprezado e o primeiro termo, o que designa a complexidade computacional propriamente dita da operação de fusão, ou seja, O_e, é que determina qual fusão possui melhor complexidade computacional.

A complexidade computacional dos OFPs aplicados é apresentada na Tabela 14, seguindo a notação acordada nesta subseção. Assim, a complexidade dos OFPs torna-se assintoticamente

Tabela 14 – Complexidade Computacional dos OFPs aplicados

OFP	**Complexidade Computacional**
Concatenação	$nk + O_t(nk)$
Soma	$nk + O_t(k)$
Máximo	$nk + O_t(k)$

equivalente pela dominância do termo nk. Ou seja, a complexidade dos operadores é dita em função da quantidade de vetores de características a serem fundidos e de suas dimensionalidades. Contudo, a saída do operador de concatenação, para valores de n muito grandes, ou seja, quando há muitos vetores a serem fundidos, acaba tornando a parcela $O_t(nk)$ relevante. Isso acaba fazendo a complexidade desse operador ser determinada pela complexidade da tarefa em questão no caso de muitas características diferentes serem fundidas. No caso deste experimento, como $n = 3$, essa parcela não se torna preponderante.

Na técnica de Kishi, Trojahn e Goularte (2019), a utilização do algoritmo de agrupamento *K-Means* torna a sua complexidade a dominante em seu processo de fusão. Assim, a complexidade dessa técnica acaba sendo assintoticamente equivalente a do algoritmo de clusterização. A complexidade computacional do *K-Means* é dada por $O(knT)$, em que k representa o número de grupos gerados pelo algoritmo; n, a quantidade de vetores de características a serem agrupados; e T, o número de iterações (ARTHUR; VASSILVITSKII, 2006). Como $O(knT)$ é assintoticamente superior a $O(nk)$, os OFPs acabam tendo uma complexidade melhor em relação a essa técnica.

Por empregar RNNs e CNNs, a complexidade computacional da técnica de Trojahn e Goularte (2021) se dá pela soma das complexidades dessas 2 arquiteturas. A complexidade de uma RNN é $O(lnd^2)$; enquanto a de uma CNN, $O(lknd^2)$, em que l representa o número de camadas; n, a dimensionalidade da entrada; d, a dimensionalidade da saída e k, o tamanho do

kernel das convoluções empregado na CNN (VASWANI *et al.*, 2017). Assim, a complexidade computacional da técnica de Trojahn e Goularte (2021) acaba sendo dominada pela complexidade da CNN utilizada. Como $O(lknd^2)$ é assintoticamente superior a $O(nk) + O_t(k)$, os OFPs acabam tendo uma complexidade melhor em relação a essa técnica.

A complexidade computacional da técnica de Pei *et al.* (2021) é determinada pela complexidade da GCN (*Graph Convolutional Network*) empregada. Já é conhecido que sua complexidade é superior à complexidade de uma CNN (PEI *et al.*, 2021). Como a complexidade de uma CNN é assintoticamente superior à dos OFPs, uma GCN também o será. Assim, os OFPs também acabam tendo uma complexidade melhor em relação a essa técnica.

CAPÍTULO

6

CONCLUSÃO

Na pesquisa desenvolvida durante este mestrado, foram propostos Operadores de Fusão Prévia, que possuem, em sua formulação, operações mais simples, comparadas às técnicas de fusão prévia do estado da arte. Com os resultados obtidos na tarefa de segmentação temporal de vídeo em cenas, pode-se concluir que eles se mostraram significativamente similares e competitivos, com resultados próximos aos de técnicas do estado da arte, dada a sua simplicidade computacional.

6.1 Contribuições

A principal contribuição deste trabalho é a aplicação de operadores de fusão prévia computacionalmente mais simples do que as atuais técnicas de fusão prévia multimodal encontradas na literatura (KISHI; TROJAHN; GOULARTE, 2019; PEI *et al.*, 2021; TROJAHN; GOULARTE, 2021). O resultado da aplicação dos operadores são representações multimodais capazes de serem aplicadas a diferentes tarefas de análise de vídeo. Neste trabalho foi explorada a segmentação temporal de vídeo em cenas como tarefa. Além disso, com um *pipeline* modular e distinguível em suas etapas, o uso de operadores torna mais fácil a separação dos estágios envolvidas no processo de análise. Com isso, pode-se analisar melhor como cada passo contribui para o resultado final, que é a avaliação da saída do algoritmo da tarefa de acordo com as métricas associadas a ela.

Com a aplicabilidade de operações mais simples, pode-se focar em outros pontos da tarefa que são passíveis de melhoria, tais como a seleção de quadros-chave, a extração de características e o algoritmo propriamente dito da tarefa. No caso específico da segmentação temporal de vídeo, como a escolha do operador ou técnica não diferiu significativamente os resultados, pode ser vantajoso utilizar os OFPs, direcionando esforços nas outras etapas do *pipeline*.

Durante a condução e execução das atividades deste mestrado, um artigo foi publicado no WebMedia 2020 (BESERRA; KISHI; GOULARTE, 2020), que apresenta e discute resultados prévios obtidos pelo uso de OFPs em outras tarefas, como a classificação de vídeo. Durante a escrita desta dissertação, um segundo artigo, intitulado *Multimodal Early Fusion Operators for Temporal Video Scene Segmentation Tasks* foi submetido ao periódico *Multimedia Tools and Applications*, que já se encontra em fase de revisão, aguardando respostas dos revisores após a primeira rodada de revisões. Além disso, o trabalho desenvolvido neste mestrado gerou como contribuição a formação de recursos humanos qualificados a nível de mestrado.

6.2 Limitações e Trabalhos Futuros

Apesar dos resultados competitivos alcançados pelos OFPs no espaço de características de médio nível, eles ainda são ultrapassados por técnicas do estado da arte mais elaboradas. Isso se deve a sua inerente simplicidade em não envolver processos de análise dos dados a serem fundidos. Esse aspecto reduz a complexidade do processo da tarefa de análise de vídeo como um todo, beneficiando pesquisadores já que os esforços antes empregados em como fundir a informação podem ser melhor utilizados em analisar o que há de ser melhorado em outros passos do *pipeline* de uma tarefa. No caso da segmentação temporal de vídeo em cenas, por exemplo, tais passos podem envolver uma melhoria do algoritmo de segmentação, o STG, ou um novo; ou a extração de novas características ou uma geração alternativa para as características de médio nível.

Por outro lado, há a falta de uma interpretação compreensível na representação da informação fundida. Como os dados de diversas modalidades não estão diretamente relacionados, uma etapa de análise se faz necessária para expor e evidenciar essa correlação intra e intermodal, como aplicado em Kishi, Trojahn e Goularte (2019). Essa etapa é ausente na aplicação dos OFPs, devido a sua inerente simplicidade. Utilizar os operadores diretamente nos dados pode reduzir a rastreabilidade para a informação original, ou seja, as representações unimodais, dificultando uma análise mais profunda da etapa de fusão.

Como trabalhos futuros, mais estudos poderiam ser realizados utilizando os OFPs em outras tarefas de análise de vídeo, analisando o impacto dos operadores em outros *datasets* e tarefas, assim contribuindo para garantir que os OFPs são de fato independente de especificidades de domínio ou tarefa. Outras modalidades também poderiam ser empregadas na etapa de fusão, como a textual, assim podendo se verificar a independência dos OFPs em relação à quantidade de características e modalidades fundidas. Outro trabalho possível seria explorar novas maneiras de detectar e expor a semântica multimodal presente nos dados, além de usar correlação, o que pode levar a uma melhoria na eficácias das tarefas.

REFERÊNCIAS

ABDEL-HAKIM, A. E.; FARAG, A. A. Csift: A sift descriptor with color invariant characteristics. In: **2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)**. [S.l.: s.n.], 2006. v. 2, p. 1978–1983. Citado nas páginas 27 e 29.

ABDU, S. A.; YOUSEF, A. H.; SALEM, A. Multimodal video sentiment analysis using deep learning approaches, a survey. **Information Fusion**, v. 76, p. 204–226, 2021. ISSN 1566-2535. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1566253521001299>. Citado na página 21.

AGGARWAL, C. C.; ZHAI, C. A survey of text classification algorithms. In: ____. **Mining Text Data**. Boston, MA: Springer US, 2012. p. 163–222. ISBN 978-1-4614-3223-4. Disponível em: <https://doi.org/10.1007/978-1-4614-3223-4_6>. Citado na página 32.

ALAM, A.; ULLAH, I.; LEE, Y. K. Video big data analytics in the cloud: A reference architecture, survey, opportunities, and open research issues. **IEEE Access**, v. 8, p. 152377–152422, 2020. Citado na página 22.

ALÍAS, F.; SOCORÓ, J.; SEVILLANO, X. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. **Applied Sciences**, MDPI AG, v. 6, n. 5, p. 143, May 2016. ISSN 2076-3417. Disponível em: <http://dx.doi.org/10.3390/app6050143>. Citado na página 31.

ARTHUR, D.; VASSILVITSKII, S. How slow is the k-means method? In: **Proceedings of the Twenty-Second Annual Symposium on Computational Geometry**. New York, NY, USA: Association for Computing Machinery, 2006. (SCG '06), p. 144–153. ISBN 1595933409. Disponível em: <https://doi.org/10.1145/1137856.1137880>. Citado na página 64.

ATREY, P. K.; HOSSAIN, M. A.; SADDIK, A. E.; KANKANHALLI, M. S. Multimodal fusion for multimedia analysis: a survey. **Multimedia Systems**, Springer Science and Business Media LLC, v. 16, n. 6, p. 345–379, abr. 2010. Disponível em: <https://doi.org/10.1007/s00530-010-0182-0>. Citado nas páginas 22, 23, 26, 34 e 35.

BARALDI, L.; GRANA, C.; CUCCHIARA, R. A deep siamese network for scene detection in broadcast videos. In: **Proceedings of the 23rd ACM International Conference on Multimedia**. New York, NY, USA: Association for Computing Machinery, 2015. (MM '15), p. 1199–1202. ISBN 9781450334594. Disponível em: <https://doi.org/10.1145/2733373.2806316>. Citado nas páginas 35, 38, 50, 53 e 54.

____. Measuring scene detection performance. In: **Pattern Recognition and Image Analysis**. Cham: Springer International Publishing, 2015. p. 395–403. ISBN 978-3-319-19390-8. Citado nas páginas 38 e 51.

BESERRA, A. A. R.; KISHI, R. M.; GOULARTE, R. Evaluating early fusion operators at mid-level feature space. In: **Proceedings of the Brazilian Symposium on Multimedia and the Web**. New York, NY, USA: Association for Computing Machinery, 2020. (WebMedia '20), p.

113–120. ISBN 9781450381963. Disponível em: <https://doi.org/10.1145/3428658.3431079>. Citado na página 68.

BROWN, M.; LOWE, D. Invariant features from interest point groups. In: **Proceedings of the British Machine Vision Conference**. [S.l.]: BMVA Press, 2002. p. 23.1–23.10. ISBN 1-901725-19-7. Citado na página 28.

CAMBRIA, E.; HUSSAIN, A. Sentic album: Content-, concept-, and context-based online personal photo management system. **Cognitive Computation**, Springer Science and Business Media LLC, v. 4, n. 4, p. 477–496, maio 2012. Disponível em: <https://doi.org/10.1007/s12559-012-9145-4>. Citado na página 33.

CETIC. **Cresce o uso de Internet durante a pandemia e número de usuários no Brasil chega a 152 milhões, é o que aponta pesquisa do Cetic.br**. 2021. Disponível em <https://bit.ly/3y2a42y>. Citado na página 21.

CHAPMAN, N.; CHAPMAN, J. **Digital multimedia**. 3rd. ed. Chichester: John Wiley, 2009. ISBN 0470512164. Citado na página 25.

CHINCHOR, N. A.; CHRISTEL, M. G.; RIBARSKY, W. Guest editors' introduction: Multimedia analytics. **IEEE Computer Graphics and Applications**, v. 30, n. 5, p. 18–19, 2010. Citado na página 22.

COUR, T.; JORDAN, C.; MILTSAKAKI, E.; TASKAR, B. Movie/script: Alignment and parsing of video and text transcription. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 5305 LNCS, n. PART 4, p. 158–171, 2008. Cited By 44. Disponível em: <https://doi.org/10.1007/978-3-540-88693-8-12>. Citado na página 26.

CSURKA, G.; DANCE, C.; FAN, L.; WILLAMOWSKI, J.; BRAY, C. Visual categorization with bags of keypoints. In: PRAGUE. **Workshop on statistical learning in computer vision, ECCV**. [S.l.], 2004. v. 1, p. 1–22. Citado nas páginas 32 e 33.

DAVIS, S.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. **IEEE Transactions on Acoustics, Speech, and Signal Processing**, v. 28, n. 4, p. 357–366, 1980. Citado na página 31.

DUSMANU, M.; ROCCO, I.; PAJDLA, T.; POLLEFEYS, M.; SIVIC, J.; TORII, A.; SATTLER, T. D2-net: A trainable CNN for joint description and detection of local features. In: **2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. IEEE, 2019. Disponível em: <https://doi.org/10.1109/cvpr.2019.00828>. Citado na página 51.

FABRO, M. D.; BÖSZÖRMENYI, L. State-of-the-art and future challenges in video scene detection: a survey. **Multimedia Systems**, Springer Science and Business Media LLC, v. 19, n. 5, p. 427–454, fev. 2013. Disponível em: <https://doi.org/10.1007/s00530-013-0306-4>. Citado nas páginas 26 e 37.

FOSS, J.; SHIRLEY, B.; MALHEIRO, B.; KEPPLINGER, S.; NIXON, L.; PHILIPP, B.; MEZARIS, V.; ULISSES, A. Datatv 2019: 1st international workshop on data-driven personalisation of television. In: **Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video**. New York, NY, USA: Association for Computing Machinery, 2019. (TVX '19), p. 286–292. ISBN 9781450360173. Disponível em: <https://doi.org/10.1145/3317697.3323349>. Citado na página 22.

FOX, M.; SCHOEFFMANN, K. The impact of dataset splits on classification performance in medical videos. In: **Proceedings of the 2022 International Conference on Multimedia Retrieval**. New York, NY, USA: Association for Computing Machinery, 2022. (ICMR '22), p. 6–10. ISBN 9781450392389. Disponível em: <https://doi.org/10.1145/3512527.3531424>. Citado na página 27.

GAONKAR, A.; CHUKKAPALLI, Y.; RAMAN, P. J.; SRIKANTH, S.; GURUGOPINATH, S. A comprehensive survey on multimodal data representation and information fusion algorithms. In: **2021 International Conference on Intelligent Technologies (CONIT)**. [S.l.]: IEEE, 2021. Citado na página 23.

GRAUMAN, K.; LEIBE, B. **Visual object recognition**. San Rafael, Calif: Morgan & Claypool Publishers, 2011. ISBN 9781598299687. Citado na página 27.

GROSS, B. M. The managing of organizations: The administrative struggle, vols. i and ii. **The ANNALS of the American Academy of Political and Social Science**, v. 360, n. 1, p. 197–198, 1965. Citado na página 22.

GÜDER, M.; ÇIÇEKLI, N. K. Multi-modal video event recognition based on association rules and decision fusion. **Multimedia Systems**, Springer Science and Business Media LLC, v. 24, n. 1, p. 55–72, fev. 2017. Disponível em: <https://doi.org/10.1007/s00530-017-0535-z>. Citado nas páginas 25 e 41.

HAN, B.; WU, W. Video scene segmentation using a novel boundary evaluation criterion and dynamic programming. In: **2011 IEEE International Conference on Multimedia and Expo**. [S.l.: s.n.], 2011. p. 1–6. Citado nas páginas 38, 39, 51 e 55.

HANJALIC, A.; LAGENDIJK, R. L.; BIEMOND, J. Automated high-level movie segmentation for advanced video-retrieval systems. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 9, n. 4, p. 580–588, 1999. Citado nas páginas 26 e 38.

HAVALDAR, P.; MEDIONI, G. **Multimedia Systems: Algorithms, Standards, and Industry Practices**. 1st. ed. Boston, MA, USA: Course Technology Press, 2009. ISBN 1418835943. Citado na página 25.

HIEMSTRA, D.; KRAAIJ, W. Evaluation of multimedia retrieval systems. In: ____. **Multimedia Retrieval**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 347–366. ISBN 978-3-540-72895-5. Disponível em: <https://doi.org/10.1007/978-3-540-72895-5_13>. Citado na página 35.

HO, K. M.; LO, K. T.; FENG, J. Multimedia streaming on the internet. In: ____. **Encyclopedia of Multimedia**. Boston, MA: Springer US, 2008. p. 614–621. ISBN 978-0-387-78414-4. Disponível em: <https://doi.org/10.1007/978-0-387-78414-4_153>. Citado na página 21.

IMRAN, A. S.; KASTRATI, Z.; SVENDSEN, T. K.; KURTI, A. Text-independent speaker id employing 2d-cnn for automatic video lecture categorization in a mooc setting. In: **2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)**. Portland, OR, USA: [s.n.], 2019. p. 273–277. Citado na página 31.

JHUO, I.-H.; YE, G.; GAO, S.; LIU, D.; JIANG, Y.-G.; LEE, D. T.; CHANG, S.-F. Discovering joint audio–visual codewords for video event detection. **Machine Vision and Applications**, Springer Science and Business Media LLC, v. 25, n. 1, p. 33–47, out. 2014. Disponível em: <https://doi.org/10.1007/s00138-013-0567-0>. Citado nas páginas 23 e 42.

JI, H.; HOOSHYAR, D.; KIM, K.; LIM, H. A semantic-based video scene segmentation using a deep neural network. **Journal of Information Science**, v. 45, n. 6, p. 833–844, 2019. Disponível em: <https://doi.org/10.1177/0165551518819964>. Citado na página 27.

JOTHILAKSHMI, S.; GUDIVADA, V. Chapter 10 - large scale data enabled evolution of spoken language research and applications. In: **Cognitive Computing: Theory and Applications**. Elsevier, 2016, (Handbook of Statistics, v. 35). p. 301 – 340. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0169716116300463>. Citado na página 31.

KENDER, J. R.; YEO, B.-L. Video scene segmentation via continuous video coherence. In: **Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)**. [S.l.: s.n.], 1998. p. 367–373. Citado na página 27.

KISHI, R. M. **Fusão de informação multimodal por detecção de correlação para tarefas de análise de vídeo**. Tese (Doutorado) — Universidade de Sao Paulo, 2020. Disponível em: <https://doi.org/10.11606/t.55.2020.tde-29072020-100439>. Citado nas páginas 34, 35, 50, 51, 52, 53, 54 e 55.

KISHI, R. M.; TROJAHN, T. H.; GOULARTE, R. Temporal video scene segmentation by fused bags-of-features. In: **Proceedings of the 24th Brazilian Symposium on Multimedia and the Web - WebMedia'18**. ACM Press, 2018. Disponível em: <https://doi.org/10.1145/3243082.3243109>. Citado nas páginas 23 e 52.

____. Correlation based feature fusion for the temporal video scene segmentation task. **Multimedia Tools and Applications**, 78, n. 11, p. 15623–15646, JUN 2019. ISSN 1380-7501. Citado nas páginas 23, 39, 42, 50, 51, 52, 53, 54, 57, 60, 64, 67 e 68.

KRAAIJ, W.; SMEATON, A. F.; OVER, P. Trecvid 2004 - an overview. In: NIST, USA. **Proceedings of the TRECVID Workshop (TRECVID 2004)**. Gaithersburg, MD, USA, 2004. Citado na página 26.

KRESS, G. **Multimodality**. [S.l.: s.n.], 2005. 179-199 p. Citado na página 34.

LAKSHMI, K. D.; VAITHIYANATHAN, V. Image registration techniques based on the scale invariant feature transform. **IETE Technical Review**, Taylor & Francis, v. 34, n. 1, p. 22–29, 2017. Disponível em: <https://doi.org/10.1080/02564602.2016.1141076>. Citado na página 30.

LEUNG, T.; MALIK, J. Representing and recognizing the visual appearance of materials using three-dimensional textons. **International Journal of Computer Vision**, Springer Science and Business Media LLC, v. 43, n. 1, p. 29–44, 2001. Disponível em: <https://doi.org/10.1023/a:1011126920638>. Citado na página 32.

LEVENE, H. Robust tests for equality of variances. **Contributions to probability and statistics. Essays in honor of Harold Hotelling**, Stanford University Press, p. 279–292, 1961. Citado na página 56.

LI, S. A review of feature detection and match algorithms for localization and mapping. **IOP Conference Series: Materials Science and Engineering**, IOP Publishing, v. 231, sep 2017. Disponível em: <https://doi.org/10.1088/1757-899X/231/1/012003>. Citado na página 28.

LIN, J.; ZHAO, Y.; LIU, C.; PU, H. Abnormal video homework automatic detection system. **Journal of Ambient Intelligence and Humanized Computing**, Springer Science and Business Media LLC, v. 12, n. 12, p. 10529–10537, jan. 2021. Disponível em: <https://doi.org/10.1007/s12652-020-02860-9>. Citado na página 31.

LOPES, B.; TROJAHN, T.; GOULARTE, R. Video scene detection by multimodal bag of features. **Journal of Information and Data Management**, v. 5, p. 1, 06 2014. Citado nas páginas 22 e 51.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. **International Journal of Computer Vision**, Springer Science and Business Media LLC, v. 60, n. 2, p. 91–110, nov. 2004. Disponível em: <https://doi.org/10.1023/b:visi.0000029664.99615.94>. Citado na página 27.

MACFARLANE, A. Knowledge organisation and its role in multimedia information retrieval. **Knowledge Organization**, v. 43, n. 3, p. 180–183, 2016. Citado na página 26.

MAHMOODZADEH, A. Human activity recognition based on deep belief network classifier and combination of local and global features. **Journal of Information Systems and Telecommunication (JIST)**, Iranian Academic Center for Education,Culture and Research, v. 9, n. 1, 2021. ISSN 2322-1437. Disponível em: <rimag.ricest.ac.ir/fa/Article/15439>. Citado na página 27.

MARTINET, J.; SAYAD, I. E. Mid-level image descriptors. In: MA, Z. (Ed.). **Intelligent Multimedia Databases and Information Retrieval: Advancing Applications and Technologies**. IGI Global, 2012. p. 46–60. Disponível em: <https://hal.archives-ouvertes.fr/hal-00730576>. Citado nas páginas 26, 27, 32 e 33.

MITROVIĆ, D.; ZEPPELZAUER, M.; BREITENEDER, C. Chapter 3 - features for content-based audio retrieval. In: **Advances in Computers: Improving the Web**. Elsevier, 2010, (Advances in Computers, v. 78). p. 71 – 150. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0065245810780037>. Citado na página 31.

MORGAN, N.; BOURLARD, H. A. Neural networks for statistical recognition of continuous speech. **Proceedings of the IEEE**, v. 83, n. 5, p. 742–772, 1995. Citado na página 31.

MÜHLING, M.; EWERTH, R.; ZHOU, J.; FREISLEBEN, B. Multimodal video concept detection via bag of auditory words and multiple kernel learning. In: **Advances in Multimedia Modeling**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 40–50. ISBN 978-3-642-27355-1. Citado nas páginas 32 e 33.

MYLONAS, P.; AVRITHIS, Y. Multimedia personalization. In: ____. **Encyclopedia of Multimedia**. Boston, MA: Springer US, 2008. p. 588–589. ISBN 978-0-387-78414-4. Disponível em: <https://doi.org/10.1007/978-0-387-78414-4_50>. Citado na página 22.

NAGASAKA, A.; TANAKA, Y. Automatic video indexing and full-video search for object appearances (abstract). **J. Inf. Process.**, Information Processing Society of Japan, JPN, v. 15, n. 2, p. 316, jan. 1992. ISSN 0387-6101. Citado na página 27.

Netflix. **Letter to Shareholders 2020**. 2022. Disponível em <http://ir.netflix.com>. Citado na página 21.

NIGAY, L.; COUTAZ, J. A design space for multimodal systems: Concurrent processing and data fusion. In: **Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems**. New York, NY, USA: Association for Computing Machinery, 1993. (CHI '93), p. 172–178. ISBN 0897915755. Disponível em: <https://doi.org/10.1145/169059.169143>. Citado na página 34.

PARK, E.; CUI, X.; KIM, W.; KIM, H. **End-to-End Fingerprints Liveness Detection using Convolutional Networks with Gram module**. 2018. Citado na página 23.

PATEL, T. A.; DABHI, V. K.; PRAJAPATI, H. B. Survey on scene classification techniques. In: **2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)**. [S.l.: s.n.], 2020. p. 452–458. Citado na página 27.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citado na página 52.

PEI, Y.; WANG, Z.; CHEN, H.; HUANG, B.; TU, W. Video scene detection based on link prediction using graph convolution network. In: **Proceedings of the 2nd ACM International Conference on Multimedia in Asia**. ACM, 2021. Disponível em: <https://doi.org/10.1145/3444685.3446293>. Citado nas páginas 39, 42, 43, 50, 57, 58, 60, 65 e 67.

PEREIRA JR., O.; FERRAZ, C. T.; GONZAGA, A. Image correspondence using a fusion of local region descriptors. In: **XIV Workshop de Visão Computacional**. [S.l.: s.n.], 2018. Citado na página 23.

POORNA, S. S.; NAIR, S. D.; NARAYAN, A.; PRASAD, V.; HIMAJA, P. S.; KAMATH, S. S.; NAIR, G. J. Bimodal emotion recognition using audio and facial features. **Journal of Computational and Theoretical Nanoscience**, v. 17, n. 1, p. 189–194, 2020. ISSN 1546-1955. Disponível em: <https://www.ingentaconnect.com/content/asp/jctn/2020/00000017/00000001/art00030>. Citado na página 31.

POUYANFAR, S.; YANG, Y.; CHEN, S.-C.; SHYU, M.-L.; IYENGAR, S. S. Multimedia big data analytics: A survey. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 51, n. 1, 2018. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3150226>. Citado nas páginas 22 e 35.

RAJARAMAN, A.; ULLMAN, J. D. Data mining. In: ____. **Mining of Massive Datasets**. [S.l.]: Cambridge University Press, 2011. p. 1–17. Citado na página 32.

RAO, A.; XU, L.; XIONG, Y.; XU, G.; HUANG, Q.; ZHOU, B.; LIN, D. A local-to-global approach to multi-modal movie scene segmentation. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2020. p. 10146–10155. Citado nas páginas 42 e 50.

RAO, K. S.; KOOLAGUDI, S. G. **Emotion Recognition using Speech Features**. Springer New York, 2013. Disponível em: <https://doi.org/10.1007/978-1-4614-5143-3>. Citado na página 31.

RAVAL, K. R.; GOYANI, M. M. A survey on event detection based video summarization for cricket. **Multimedia Tools and Applications**, Springer Science and Business Media LLC, v. 81, n. 20, p. 29253–29281, abr. 2022. Disponível em: <https://doi.org/10.1007/s11042-022-12834-y>. Citado na página 21.

RIJSBERGEN, C. J. V. **Information Retrieval**. 2nd. ed. USA: Butterworth-Heinemann, 1979. ISBN 0408709294. Citado na página 37.

ROTMAN, D.; PORAT, D.; ASHOUR, G. Robust video scene detection using multimodal fusion of optimally grouped features. In: . [S.l.: s.n.], 2017. v. 2017-January, p. 1–6. Citado nas páginas 35, 36 e 51.

RUI, Y.; HUANG, T.; MEHROTRA, S. Constructing table-of-content for videos. **Multimedia Systems**, v. 7, n. 5, p. 359–368, 1999. Disponível em: <https://dor.org/10.1007/s005300050138>. Citado na página 26.

SANDE, K. E. A. van de; GEVERS, T.; SNOEK, C. G. M. Empowering visual categorization with the gpu. **IEEE Transactions on Multimedia**, v. 13, n. 1, p. 60–70, 2011. Disponível em: <http://www.science.uva.nl/research/publications/2011/vandeSandeITM2011>. Citado na página 51.

SARACENO, C.; LEONARDI, R. Audio as a support to scene change detection and characterization of video sequences. In: **1997 IEEE International Conference on Acoustics, Speech, and Signal Processing**. Munich: [s.n.], 1997. v. 4, p. 2597–2600. Citado na página 26.

SEN, S.; DUTTA, A.; DEY, N. **Audio Processing and Speech Recognition**. Springer Singapore, 2019. Disponível em: <https://doi.org/10.1007/978-981-13-6098-5>. Citado na página 51.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, v. 52, n. 3-4, p. 591–611, 12 1965. ISSN 0006-3444. Disponível em: <https://doi.org/10.1093/biomet/52.3-4.591>. Citado na página 56.

SHARAFI, M.; YAZDCHI, M.; RASTI, R.; NASIMI, F. A novel spatio-temporal convolutional neural framework for multimodal emotion recognition. **Biomedical Signal Processing and Control**, v. 78, p. 103970, 2022. ISSN 1746-8094. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1746809422004694>. Citado na página 31.

SHARMA, D.; ALI, I. A modified MFCC feature extraction technique for robust speaker recognition. In: **2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)**. IEEE, 2015. Disponível em: <https://doi.org/10.1109/icacci.2015.7275749>. Citado na página 31.

SIDIROPOULOS, P.; MEZARIS, V.; KOMPATSIARIS, I.; MEINEDO, H.; BUGALHO, M.; TRANCOSO, I. Temporal video segmentation to scenes using high-level audiovisual features. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 21, n. 8, p. 1163–1177, 2011. Citado nas páginas 22 e 27.

SIDIROPOULOS, P.; MEZARIS, V.; KOMPATSIARIS, I.; KITTLER, J. Differential edit distance: A metric for scene segmentation evaluation. **IEEE Transactions on Circuits and Systems for Video Technology**, v. 22, n. 6, p. 904–914, 2012. Citado na página 38.

SINGH, L.; CHETTY, G. A comparative study of recognition of speech using improved MFCC algorithms and rasta filters. In: **Information Systems, Technology and Management**. Springer Berlin Heidelberg, 2012. p. 304–314. Disponível em: <https://doi.org/10.1007/978-3-642-29166-1_27>. Citado na página 31.

SMEULDERS, A. W. M.; WORRING, M.; SANTINI, S.; GUPTA, A.; JAIN, R. Content-based image retrieval at the end of the early years. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 12, p. 1349–1380, 2000. Citado na página 26.

SNOEK, C. G. M.; WORRING, M. A review on multimodal video indexing. In: **Proceedings. IEEE International Conference on Multimedia and Expo**. [S.l.: s.n.], 2002. v. 2, p. 21–24. Citado nas páginas 22 e 34.

____. Concept-based video retrieval. **Foundations and Trends® in Information Retrieval**, v. 2, n. 4, p. 215–322, 2009. ISSN 1554-0669. Disponível em: <http://dx.doi.org/10.1561/1500000014>. Citado na página 26.

SNOEK, C. G. M.; WORRING, M.; SMEULDERS, A. W. M. Early versus late fusion in semantic video analysis. In: **Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA'05**. ACM Press, 2005. Disponível em: <https://doi.org/10.1145/1101149.1101236>. Citado na página 41.

STEVENS, S.; VOLKMANN, J.; NEWMAN, E. A scale for the measurement of the psychological magnitude pitch. **Journal of the Acoustical Society of America**, v. 8, n. 3, p. 185–190, 1937. Citado na página 31.

SUHASINI, P. S.; KRISHNA, K. S. R.; KRISHNA, I. V. M. Content based image retrieval based on different global and local color histogram methods: A survey. **Journal of The Institution of Engineers (India): Series B**, Springer Science and Business Media LLC, v. 98, n. 1, p. 129–135, jun. 2016. Disponível em: <https://doi.org/10.1007/s40031-016-0223-y>. Citado na página 27.

SUMALAKSHMI, C. H.; VASUKI, P. Fused deep learning based facial expression recognition of students in online learning mode. **Concurrency and Computation: Practice and Experience**, v. 34, n. 21, p. 7137, 2022. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.7137>. Citado na página 21.

SUNDARAM, H.; CHANG, S.-F. Computable scenes and structures in films. **IEEE Transactions on Multimedia**, v. 4, n. 4, p. 482–491, 2002. Citado na página 26.

TROJAHN, T. H.; GOULARTE, R. Temporal video scene segmentation using deep-learning. **Multimedia Tools and Applications**, Springer Science and Business Media LLC, v. 80, n. 12, p. 17487–17513, fev. 2021. Disponível em: <https://doi.org/10.1007/s11042-020-10450-2>. Citado nas páginas 42, 43, 57, 58, 60, 64, 65 e 67.

TROJAHN, T. H.; KISHI, R. M.; GOULARTE, R. A new multimodal deep-learning model to video scene segmentation. In: **Proceedings of the 24th Brazilian Symposium on Multimedia and the Web - WebMedia '18**. ACM Press, 2018. Disponível em: <https://doi.org/10.1145/3243082.3243108>. Citado na página 23.

TUYTELAARS, T.; MIKOLAJCZYK, K. Local invariant feature detectors: A survey. **Foundations and Trends® in Computer Graphics and Vision**, v. 3, n. 3, p. 177–280, 2008. ISSN 1572-2740. Disponível em: <http://dx.doi.org/10.1561/0600000017>. Citado na página 27.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention is all you need. In: . [S.l.: s.n.], 2017. v. 2017-December, p. 5999–6009. Citado na página 65.

VEMBU, A.; NATARAJAN, P.; WU, S.; PRASAD, R.; NATARAJAN, P. Graph based multimodal word clustering for video event detection. In: **2013 IEEE International Conference on Acoustics, Speech and Signal Processing**. IEEE, 2013. Disponível em: <https://doi.org/10.1109/icassp.2013.6638342>. Citado na página 42.

VENDRIG, J.; WORRING, M. Systematic evaluation of logical story unit segmentation. **IEEE Transactions on Multimedia**, Institute of Electrical and Electronics Engineers (IEEE), v. 4, n. 4, p. 492–499, dez. 2002. Disponível em: <https://doi.org/10.1109/tmm.2002.802021>. Citado nas páginas 38 e 55.

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J.; van der Walt, S. J.; BRETT, M.; WILSON, J.; MILLMAN, K. J.; MAYOROV, N.; NELSON, A. R. J.; JONES, E.; KERN, R.; LARSON, E.; CAREY, C. J.; POLAT, İ.; FENG, Y.; MOORE, E. W.; VanderPlas, J.; LAXALDE, D.; PERKTOLD, J.; CIMRMAN, R.; HENRIKSEN, I.; QUINTERO, E. A.; HARRIS, C. R.; ARCHIBALD, A. M.; RIBEIRO, A. H.; PEDREGOSA, F.; van Mulbregt, P.; SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. **Nature Methods**, v. 17, p. 261–272, 2020. Citado na página 56.

WANG, K.; BICHOT, C.-E.; LI, Y.; LI, B. Local binary circumferential and radial derivative pattern for texture classification. **Pattern Recognition**, v. 67, p. 213 – 229, 2017. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320317300407>. Citado nas páginas 23 e 41.

WANG, Z.; WANG, E.; WANG, S.; DING, Q. Multimodal biometric system using face-iris fusion feature. **JCP**, v. 6, p. 931–938, 2011. Citado na página 41.

WEI, Y.; BHANDARKAR, S. M.; LI, K. Video personalization in resource-constrained multimedia environments. In: **Proceedings of the 15th ACM International Conference on Multimedia**. New York, NY, USA: Association for Computing Machinery, 2007. (MM '07), p. 902–911. ISBN 9781595937025. Disponível em: <https://doi.org/10.1145/1291233.1291436>. Citado na página 21.

WELCH, B. L. THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED. **Biometrika**, v. 34, n. 1-2, p. 28–35, 01 1947. ISSN 0006-3444. Disponível em: <https://doi.org/10.1093/biomet/34.1-2.28>. Citado na página 56.

YANG, J.; LANG, L.; SONG, S. A study of data-driven enterprise human resource management model. **Discrete Dynamics in Nature and Society**, Hindawi Limited, v. 2021, p. 1–11, nov. 2021. Citado na página 42.

YANG, L.; WANG, Y.; DUNNE, D.; SOBOLEV, M.; NAAMAN, M.; ESTRIN, D. More than just words: Modeling non-textual characteristics of podcasts. In: **Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2019. (WSDM '19), p. 276–284. ISBN 9781450359405. Disponível em: <https://doi.org/10.1145/3289600.3290993>. Citado na página 31.

YEUNG, M.; YEO, B.-L.; LIU, B. Segmentation of video by clustering and graph analysis. **Computer Vision and Image Understanding**, v. 71, n. 1, p. 94 – 109, 1998. ISSN 1077-3142. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1077314297906287>. Citado nas páginas 53 e 54.

Youtube. **Youtube para a imprensa**. 2021. Disponível em <https://www.youtube.com/about/press/>. Citado na página 21.

ZHANG, H.; LOW, C. Y.; SMOLIAR, S. W. Video parsing and browsing using compressed data. **Multimedia Tools and Applications**, Springer Science and Business Media LLC, v. 1, n. 1, p. 89–111, mar. 1995. Disponível em: <https://doi.org/10.1007/bf01261227>. Citado na página 26.

ZHANG, X.; ZHANG, H.; ZHANG, Y.; YANG, Y.; WANG, M.; LUAN, H.; LI, J.; CHUA, T.-S. Deep fusion of multiple semantic cues for complex event recognition. **IEEE Transactions on Image Processing**, Institute of Electrical and Electronics Engineers (IEEE), v. 25, n. 3, p. 1033–1046, mar. 2016. Disponível em: <https://doi.org/10.1109/tip.2015.2511585>. Citado na página 41.


ICMC
USP
SÃO CARLOS