UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Um método de segmentação de vídeo em cenas baseado em aprendizagem profunda

Tiago Henrique Trojahn

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

ICMC USP SÃO CARLOS

Tiago Henrique Trojahn

Um método de segmentação de vídeo em cenas baseado em aprendizagem profunda

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor 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
Agosto de 2019

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

Trojahn, Tiago Henrique
T845m Um método de segmentação de vídeo em cenas baseado em aprendizagem profunda / Tiago Henrique Trojahn; orientador Rudinei Goularte. -- São Carlos, 2019.
129 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2019.

1. Aprendizagem profunda. 2. Segmentação em cenas. 3. Multimodalidade. 4. Fusão Multimodal. 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

Tiago Henrique Trojahn

A video scene segmentation method based on deep learning

Thesis submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Rudinei Goularte

USP – São Carlos
August 2019

Este trabalho é dedicado aos meus pais que,
mesmo fisicamente distantes,
sempre estiveram por perto.

AGRADECIMENTOS

Agradeço aos meus pais, Ingrid e Neldo, que sempre me incentivaram e apoiaram nos altos e baixos durante a realização tanto do mestrado como deste doutorado. Meus eternos agradecimentos.

Um muito obrigado a meus colegas e amigos do laboratório Intermídia que, diariamente, colaboraram mesmo que indiretamente para a realização deste projeto. Agradeço em especial ao Humberto, Márcio, Juliano e Rodrigo pela parceria nas festas, barzinhos, cinemas e eventuais jogatinas de final-de-semana. Thank you!

Meus agradecimentos ao Instituto de Educação, Ciência e Tecnologia de São Paulo (IFSP, campus São Carlos) pela oportunidade de poder me dedicar exclusivamente ao doutorado durante boa parte de sua duração. Agradeço em especial aos meus colegas de trabalho Elis Cristina e Jorge Cutigi pelo auxílio e companhia nos primeiros anos no instituto. Obrigado!

Um agradecimento em especial a Rodrigo M. Kishi que me auxiliou diretamente em várias etapas durante o desenvolvimento deste projeto, seja com dicas de artigos relevantes, sugestões de abordagens, implementações de ferramentas diversas, artigos escritos e bases de dados encontradas. Muito obrigado!

Não poderia deixar de agradecer ao meu orientador, Rudinei Goularte, por toda a paciência, dedicação e orientação dada tanto durante o mestrado como também no doutorado.

Por fim, agradeço também a Deus por permitir que tudo isso ocorresse.

RESUMO

TROJAHN, T. H. **Um método de segmentação de vídeo em cenas baseado em aprendizagem profunda**. 2019. 129 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

A segmentação automática de vídeo em cenas é um problema atual e relevante dado sua aplicação em diversos serviços ligado à área de multimídia. Dentre as diferentes técnicas reportadas pela literatura, as multimodais são consideradas mais promissoras, dado a capacidade de extrair informações de diferentes mídias de maneira potencialmente complementar, possibilitando obter segmentações mais significativas. Ao usar informações de diferentes naturezas, tais técnicas enfrentam dificuldades para modelar e obter uma representação combinada das informações ou com elevado custo ao processar cada fonte de informação individualmente. Encontrar uma combinação adequada de informação que aumente a eficácia da segmentação a um custo computacional relativamente baixo torna-se um desafio. Paralelamente, abordagens baseadas em Aprendizagem Profunda mostraram-se eficazes em uma ampla gama de tarefas, incluindo classificação de imagens e vídeo. Técnicas baseadas em Aprendizagem Profunda, como as Redes Neurais Convolucionais (CNNs), têm alcançado resultados impressionantes em tarefas relacionadas por conseguirem extrair padrões significativos dos dados, incluindo multimodais. Contudo, CNNs não podem aprender adequadamente os relacionamentos entre dados que estão temporalmente distribuídos entre as tomadas de uma mesma cena. Isto pode tornar a rede incapaz de segmentar corretamente cenas cujas características mudam entre tomadas. Por outro lado, Redes Neurais Recorrentes (RNNs) têm sido empregadas com sucesso em processamento textual, pois foram projetadas para analisar sequências de dados de tamanho variável e podem melhor explorar as relações temporais entre as características de tomadas relacionadas, potencialmente aumentando a eficácia da segmentação em cenas. Há uma carência de métodos de segmentação multimodais que explorem Aprendizagem Profunda. Assim, este trabalho de doutorado propõe um método automático de segmentação de vídeo em cenas que modela o problema de segmentação como um problema de classificação. O método conta com um modelo que combina o potencial de extração de padrões das CNNs com o processamento de sequencias das RNNs. O modelo proposto elimina a dificuldade de modelar representações multimodais das diferentes informações de entrada além de permitir instanciar diferentes abordagens para fusão multimodal (antecipada ou tardia). Tal método foi avaliado na tarefa de segmentação em cenas utilizando uma base de vídeos pública, comparando os resultados obtidos com os resultados de técnicas em estado-da-arte usando diferentes abordagens. Os resultados mostram um avanço significativo na eficácia obtida.

Palavras-chave: Aprendizagem profunda, segmentação em cenas, multimodalidade, fusão multimodal.

ABSTRACT

TROJAHN, T. H. **A video scene segmentation method based on deep learning**. 2019. 129 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

Automatic video scene segmentation is a current and relevant problem given its application in various services related to multimedia. Among the different techniques reported in the literature, the multimodal ones are considered more promising, given the ability to extract information from different media in a potentially complementary way, allowing for more significant segmentations. By processing information of different natures, such techniques faces difficulties on modeling and obtaining a combined representation of information and cost problems when processing each source of information individually. Finding a suitable combination of information that increases the effectiveness of segmentation at a relatively low computational cost becomes a challenge. At the same time, approaches based on Deep Learning have proven effective on a wide range of tasks, including classification of images and video. Techniques based on Deep Learning, such as Convolutional Neural Networks (CNNs), have achieved impressive results in related tasks by being able to extract significant patterns from data, including multimodal data. However, CNNs can not properly learn the relationships between data temporarily distributed among the shots of the same scene. This can lead the network to become unable to properly segment scenes whose characteristics change among shots. On the other hand, Recurrent Neural Networks (RNNs) have been successfully employed in textual processing since they are designed to analyze variable-length data sequences and can be developed to better explore the temporal relationships between low-level characteristics of related shots, potentially increasing the effectiveness of scene segmentation. There is a lack of multimodal segmentation methods exploring Deep Learning. Thus, this thesis proposes an automatic method for video scene segmentation that models the problem of segmentation as a classification problem. The method relies on a model developed to combine the potential for extracting patterns from CNNs with the potential for sequence processing of the RNNs. The proposed model, different from related works, eliminates the difficulty of modeling multimodal representations of the different input information, besides allowing to instantiate different approaches for multimodal (early or late) fusion. This method was evaluated in the scene segmentation task using a public video database, comparing the results obtained with the results of state-of-the-art techniques using different approaches. The results show a significant advance in the efficiency obtained.

Keywords: Deep learning, scene segmentation, multimodality, multimodal fusion.

LISTA DE ILUSTRAÇÕES

LISTA DE ALGORITMOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BCE	Binary Cross Entropy
BoF	Bag of Features
BoVW	Bag of Visual Words
BoW	Bag of Words
BPTT	Backpropagation Through Time
BSC	Backward Shot Coherence
CCH	Contrast Context Histogram
CNN	Convolutional Neural Network
CSIFT	Color SIFT
DED	Differential Edit Distance
DFT	Discrete Fourier Transform
DoG	Difference of Gaussian
DTW	Dynamic Time Warping
FFT	Fast Fourier Transform
GRU	Gated Recurrent Unit
ILSVRC	ImageNet Large Scale Visual Recognition Competition
LFCC	Linear Frequency Cepstral Coefficient
LPC	Linear Predictor Coefficient
LPCC	Linear Predictive Cepstral Coefficients
LReLU	Leaky Rectified Linear Unit
LSTM	Long Short Term Memory
MFCC	Mel-Frequency Cepstrum Coefficients
MPEG	Moving Picture Expert Group
NW	Needleman-Wunsch
OVSD	Open Video Scene Detection
PReLU	Parametric Rectified Linear Unit
RANSAC	Random Sample Consensus
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SDN	Siamese Deep Network
SGD	Stochastic Gradient Descendent
ZCR	Zero Crossing Rate

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

O objetivo deste capítulo é introduzir o trabalho de doutorado desenvolvido. Para tanto o capítulo descreve, na Seção 1.1, uma contextualização da segmentação em vídeo e a abordagem multimodal para a segmentação em cenas, a Seção 1.2 apresenta algumas lacunas presentes no estado da arte que motivam a realização deste trabalho, a Seção 1.3 define o objetivo deste trabalho e, por fim, a Seção 1.4 descreve a organização desta tese.

1.1 Contexto

O barateamento de dispositivos capazes de produzir, armazenar e reproduzir vídeo, aliados a popularização de redes de alta velocidade, faz com que vídeo seja uma das mídias mais populares atualmente. Diversos serviços relacionados a vídeo comprovam tal interesse. Por exemplo, o YouTube[1], um dos mais conhecidos serviços gratuitos de envio e reprodução de vídeo, reporta[2] possuir quase dois bilhões de usuários logados por mês assistindo mais de um bilhão de horas de vídeo em alguma de suas 91 versões em 80 línguas distintas. Ainda, o Netflix[3], um serviço pago que disponibiliza um amplo número de filmes, documentários, seriados e similares, reporta[4] possuir mais de 130 milhões de assinantes em 190 países do mundo.

Tal volume de vídeos reflete na piora do problema conhecido como sobrecarga da informação (do inglês *information overload*), um termo popularizado por Toffler (1984), no qual o usuário enfrenta dificuldades e até desconforto devido a incapacidade de filtrar o imenso volume de conteúdo em busca de algo de seu interesse. A indexação, sumarização e recomendação multimídia (MANZATO, 2011; ADOMAVICIUS; TUZHILIN, 2005) são exemplos de serviços que procuram amenizar tal problema, facilitando ao usuário a obtenção de conteúdos de interesse,

[1] <https://www.youtube.com>
[2] <https://www.youtube.com/intl/en-GB/yt/about/press/>
[3] <https://www.netflix.com>
[4] <https://media.netflix.com/pt_br/about-netflix>

reduzindo a sobrecarga da informação.

Para serem efetivos tais serviços necessitam extrair dos vídeos uma série de diferentes dados e informações, o que tende a requerer alto custo computacional e de tempo devido ao expressivo volume de dados. Como processar um vídeo todo é geralmente impraticável, primeiramente o vídeo é segmentado em trechos menores, mais administráveis computacionalmente. A esses trechos ou segmentos é possível, então, aplicar técnicas de Indexação Multimídia (BRUNELLI; MICH; MODENA, 1999) que extraem representações mais compactas dos diferentes dados e informações presentes no conteúdo multimídia do vídeo, reduzindo o volume. A Indexação Multimídia, é importante notar, pode também ser aplicada como um processo auxiliar da segmentação.

Entre as técnicas automáticas de segmentação de vídeo possíveis se destaca a chamada segmentação temporal hierárquica, na qual um vídeo é segmentado nas partes que o compõe tal como quadros (do inglês *frames*), tomadas (do inglês *shots*) e cenas (do inglês *scenes*). A segmentação em quadros é um problema trivial dado que um vídeo, em última análise, é uma sequência de quadros Blanken *et al.* (2007) (Seção 2.1). Por sua vez, a segmentação em tomadas (Seção 2.1) é considerada um problema essencialmente resolvido (FABRO; BÖSZÖRMENYI, 2013), com técnicas robustas e de alta eficácia na maioria dos casos (SMEATON; OVER; DOHERTY, 2010; SIDIROPOULOS *et al.*, 2012). A segmentação em cenas é uma área de pesquisa ativa com diversos desafios, desde a dificuldade de se definir apropriadamente o conceito subjetivo de "cena" (FABRO; BÖSZÖRMENYI, 2013), até o como se obter o contexto ou assunto de um dado vídeo, consequência da chamada lacuna semântica (do inglês *semantic gap*) (SMEULDERS *et al.*, 2000; PROTASOV *et al.*, 2018).

O interesse neste trabalho está na segmentação automática de vídeo em cenas. Uma cena de vídeo pode ser definida, segundo Liang *et al.* (2009), Baber, Afzulpurkar e Bakhtyar (2011), Chasanis, Likas e Galatsanos (2009), Wang *et al.* (2006) e Lopes, Trojahn e Goularte (2014), como uma sequência de tomadas semanticamente relacionadas. Em geral, a relação semântica entre tomadas consecutivas se estabelece quando as mesmas tratam de um mesmo assunto. Assim, uma mudança de assunto, com consequente mudança de semântica, implica em uma mudança ou transição de cena.

O processo de segmentação em cenas geralmente adotado pela literatura, ilustrado na Figura 1, consiste em três etapas principais: o Pré-processamento, Segmentação e o Pós-processamento. Na etapa de Pré-processamento, as informações relevantes do vídeo são selecionadas e extraídas para que possam ser, já na etapa de Segmentação, analisadas para identificar, classificar ou determinar suas respectivas cenas. A segmentação em cenas resultante pode ser posteriormente analisada e refinada na etapa de Pós-processamento. O foco dos pesquisadores da área é a etapa de Segmentação, no qual buscam desenvolver abordagens capazes de identificar o contexto e o relacionamento semântico das informações obtidas na etapa de pré-processamento. Contudo, é importante ressalvar que a etapa de Pré-processamento exige dos pesquisadores

grande atenção e esforço para modelar e desenvolver adequadamente a extração das informações (BLANKEN *et al.*, 2007; Jesus; Araújo; Canuto, 2016).

Figura 1 – Etapas de processamento geralmente adotadas para a segmentação em cenas.

Fonte: Elaborada pelo autor.

As técnicas de segmentação reportadas na literatura podem ser classificadas como unimodais ou multimodais (SHEN; DEMARTY; DUONG, 2017; VUKOTIC; RAYMOND; GRAVIER, 2018; SIDIROPOULOS *et al.*, 2011). O *modo* se refere a um dos possíveis tipos de fluxos de dados presentes em um vídeo, tal como o visual e o aural, por exemplo. Abordagens unimodais, assim, utilizam apenas uma modalidade para detectar as transições de cenas, enquanto que técnicas multimodais fazem uso de pelo menos duas modalidades diferentes. Assim, a multimodalidade, quando aplicada à segmentação em cena, possibilita que informações de diferentes fontes possam ser usadas de maneira complementar, com o objetivo de melhor identificar o contexto ou semântica de um trecho de vídeo, como uma cena.

Técnicas unimodais possuem como limitação a incapacidade de detectar a mudança do assunto (cena) caso a mesma ocorra em uma modalidade que não seja a analisada. Um exemplo de tal caso são vídeos do domínio de "entrevistas", no qual o conteúdo visual, a modalidade usualmente mais utilizada (FABRO; BÖSZÖRMENYI, 2013), não apresenta qualquer mudança significativa mesmo com a mudança do assunto. Devido a essa particularidade, técnicas unimodais são limitadas a determinados domínios de vídeos.

Por outro lado, técnicas multimodais podem obter informações de maior nível semântico ao usar diversas modalidades que, aplicadas de modo potencialmente complementar, as tornam aptas a reconhecer uma maior gama de transições de cena. Por tal motivo, técnicas recentes reportadas na literatura, especialmente as sem um domínio de vídeo definido, utilizam a abordagem multimodal (SIDIROPOULOS *et al.*, 2011; LOPES; TROJAHN; GOULARTE, 2014; BARALDI; GRANA; CUCCHIARA, 2015a; BARALDI; GRANA; CUCCHIARA, 2017). Uma análise de técnicas de segmentação em cenas relacionadas é apresentada no Capítulo 3.

1.2 Motivação

Técnicas multimodais necessitam, geralmente, unir as informações provenientes de diferentes modalidades em uma única representação, em um processo chamado de fusão multimodal (ATREY; MADDAGE; KANKANHALLI, 2006). A fusão multimodal pode ocorrer em dois momentos distintos: antes do processo de segmentação em cenas (fusão antecipada), fundindo

características de diferentes modalidades em uma única representação, ou após processos de segmentações aplicados em cada modalidade individual (fusão tardia), fundindo as decisões individuais em uma segmentação única. Uma descrição detalhada das diferentes abordagens de fusão é apresentada na Seção 2.5.

Durante a última década, os pesquisadores contemplam qual a abordagem de fusão multimodal (antecipada ou tardia) é mais adequada para a tarefa de segmentação em cenas. Uma conclusão obtida é que a fusão antecipada é de maior dificuldade de modelagem quando comparada a fusão tardia (FABRO; BÖSZÖRMENYI, 2013) devido a diferentes particularidades de cada modalidade. É constatado que, porém, quando bons modelos puderam ser desenvolvidos, apresentaram eficácia bastante significativa (SNOEK; WORRING; SMEULDERS, 2005; ATREY *et al.*, 2010).

As abordagens antecipadas chamam a atenção devido a potencialmente demandarem menor custo computacional (ATREY *et al.*, 2010), contudo, as particularidades de cada modalidade levam a modelagens nas quais as representações combinadas das informações extraídas possuem alto grau de acoplamento com a técnica de segmentação em si. Isso dificulta explorar diferentes combinações de informações (cor, textura, forma, timbre, intensidade sonora, frequência de termos, etc.), nas suas mais variadas representações, a fim de melhorar tanto o custo computacional quanto a eficácia da segmentação em cenas.

Em paralelo, técnicas de Aprendizagem de Profunda (do inglês *Deep Learning*) obtiveram resultados expressivos em áreas relacionadas, como reconhecimento de faces (SCHROFF; KALENICHENKO; PHILBIN, 2015), detecção de objetos (HE *et al.*, 2015) e tradução de textos (CHO *et al.*, 2014). Redes neurais baseadas em Aprendizagem Profunda, tais como redes neurais convolucionais (do inglês *Convolutional Neural Networks* - CNN) (WIATOWSKI; BOLCSKEI, 2018) e redes neurais recorrentes (do inglês *Recurrent Neural Network* - RNN) (RUMELHART; HINTON; WILLIAMS, 1988; ELMAN, 1990), são especialmente efetivas para a classificação e reconhecimento de padrões dos dados de entrada, tornando-as interessantes para a segmentação em cenas.

Dentre suas vantagens, as redes neurais convolucionais podem facilitar a etapa de pré-processamento da segmentação em cenas. Sua principal funcionalidade é a de aprender padrões relevantes e altamente significativos das informações de entrada podendo, potencialmente, gerar uma representação compacta e homogênea de cada tomada, independentemente do tipo de modalidade analisada. Porém, conforme mencionado por Lipton, Berkowitz e Elkan (2015), uma importante limitação das redes convolucionais é que, após a sua execução, o estado anterior da rede é perdido, algo inaceitável em problemas no qual os dados possuem relacionamento temporal. Assim, uma rede convolucional não se aproveita das informações temporais presentes em tomadas de uma mesma cena, algo que pode prejudicar a qualidade da segmentação em cenas em casos no qual o assunto muda conforme o tempo. As redes neurais recorrentes, por outro lado, são um candidato natural à tarefa de segmentação em cenas. Tais redes foram espe-

cialmente desenvolvidas para o tratamento de dados sequenciais (GOODFELLOW; BENGIO; COURVILLE, 2016), portanto adequadas para processar uma sequência de quadros (tomada) ou mesmo de tomadas (cena). Ao unir as capacidades de extração de informações relevantes das redes convolucionais e o processamento de dados sequenciais das redes recorrentes, pode ser possível identificar a semântica latente das tomadas, agrupando-as em cenas de alto nível.

Apesar de suas potencialidades, até onde se tem conhecimento, abordagens de Aprendizagem Profunda foram pouco exploradas em tarefas de indexação e segmentação de vídeo. As técnicas propostas por Baraldi, Grana e Cucchiara (2015a) e Baraldi, Grana e Cucchiara (2017) utilizam uma associação de diferentes tipos de redes neurais, incluindo convolucionais, para estimar a dissimilaridade entre duas tomadas para agrupá-las em cenas. Tais técnicas e outras reportadas na literatura (discutido no Capítulo 3), porém, não abordam adequadamente as questões relativas ao forte acoplamento entre representação multimodal de características e técnicas de segmentação, assim como às relativas a uma flexibilização no uso de diferentes abordagens de fusão. Em geral pode-se dizer que isso é reflexo das arquiteturas adotadas para realizar a fusão e segmentação, que foram projetadas com base em um conjunto limitado e especifico de dados de entrada. Desse modo é possível notar uma carência de modelos e métodos que: a) deem suporte a utilização de fusão multimodal tanto antecipada quanto tardia; b) forneçam independência das características de entrada, permitindo utilizar um conjunto de tamanho e modalidades variados; c) eliminem a dificuldade de modelar representações multimodais das diferentes características de entrada.

Outro motivador deste trabalho é sua alta relevância, potencialmente beneficiando um amplo número de pessoas em atividades relacionadas a vídeo digital. A segmentação em cenas realizada neste trabalho pode auxiliar, por exemplo, o desenvolvimento e aperfeiçoamento de tarefas correlatas de vídeo, como sumarização personalizada do vídeo de acordo com o perfil do usuário. Grandes empresas com enormes repositórios de vídeos podem aplicar o método aqui proposto para obter uma segmentação parcial/total de seus repositórios de maneira a obter uma indexação mais concisa e útil de seu acervo visual, facilitando o acesso a um conteúdo de interesse. Grandes provedores de vídeo digital via internet podem, ainda, utilizar o método para auxiliar técnicas de rotulagem automática, auxiliando a indexação de seus conteúdos e a sua recomendação de acordo com os padrões de acessos dos usuários. Além disso, pesquisadores podem empregar o método aqui proposto para outras tarefas de vídeo, como classificação de vídeo ou segmentações em diferentes níveis (como capítulos ou video-aulas, por exemplo). Tarefas diversas de vídeos podem ser facilitadas ou até mesmo tornarem-se possíveis graças a segmentação em cenas, como controle parental automático de vídeo (classificação etária), classificação e rotulagem automática, auxílio a criação de conteúdo e edição de vídeo, sumarização automática de vídeos ao vivo, detecção de eventos anômalos em vídeos de segurança, entre outras.

1.3 Objetivos

Este trabalho tem por objetivo propor um método automático de segmentação temporal em cenas que auxilie na capacidade de análise semântica na etapa de Segmentação, melhorando a eficácia quando comparado com abordagens em estado da arte reportadas na literatura.

Outro objetivo é projetar um método que consiste em duas partes: 1) modelar o problema da segmentação em cenas como um problema de classificação de tomadas; 2) um modelo baseado em Aprendizagem Profunda para realizar as etapas de pré-processamento e de segmentação em um processo de segmentação temporal de vídeo em cenas. O modelo, por sua vez, consiste em uma combinação de redes convolucionais e recorrentes profundas que podem ser facilmente organizadas em duas arquiteturas diferentes, uma contemplando fusão antecipada e outra contemplando fusão tardia. Isso reduz as diferenças de nível de dificuldade de modelagem entre as duas abordagens de fusão, permitindo ao pesquisador optar por qualquer das abordagens mantendo o mesmo conjunto de características de entrada e o mesmo algoritmo de segmentação.

As redes convolucionais são capazes de extrair padrões significativos de modalidades diferentes, como visual ou aural, oferecendo uma representação homogênea para cada tomada, independentemente das particularidades de cada característica de entrada. Isso ajuda a tornar o modelo flexível quanto as características, possibilitando ao pesquisador encontrar a melhor combinação delas para melhorar a eficácia da segmentação, além de facilitar a etapa de pré-processamento em si. Já as redes recorrentes, por sua vez, podem ser projetadas de modo a analisar informações de qualquer natureza, buscando encontrar padrões temporais que auxiliem a identificação do correlacionamento semântico entre tomadas adjacentes e, consequentemente, das transições de cenas em um vídeo.

1.4 Organização do trabalho

No Capítulo 2 são descritos os conceitos relacionados essenciais para a compreensão desta tese, tal como a definição de vídeo digital, sua estrutura hierárquica, o processo de extração de características e o processo de segmentação em cenas. São descritos ainda a Aprendizagem Profunda, especificamente redes convolucionais e recorrentes, além da fusão de modalidades e questões relacionadas a avaliação da eficácia de técnicas de segmentação em cena, como base de dados e métricas amplamente utilizadas.

Por sua vez, no Capítulo 3 é descrita uma discussão de diversas técnicas de segmentação de vídeo em cenas relacionados a este trabalho reportados na literatura, tanto seminais como as consideradas em estado da arte.

Já no Capítulo 4 são apresentados o modelo proposto nesta tese, incluindo a representação da segmentação de vídeo em cenas como um problema de classificação, a extração de características multimodais do vídeo e sua fusão multimodal, por meio da abordagem anteci-

pada ou tardia, usando uma arquitetura de rede profunda composta por redes convolucionais e recorrentes.

No Capítulo 5 é apresentado o treinamento das redes neurais empregado, além de uma avaliação de eficácia da implementação do modelo proposto para o problema da segmentação em cenas, utilizando uma base de vídeos pública e adequada para a tarefa de segmentação em cenas. Considerações de eficiência relativas ao tempo necessário para o treinamento das redes neurais desenvolvidas e a obtenção da segmentação em cenas em si são discutidas. Por fim, uma comparação entre os resultados de eficácia obtidos e os reportados em técnicas relacionadas em estado da arte também é descrita.

Por fim, no Capítulo 6 são discutidas as conclusões obtidas deste trabalho e possíveis direcionamentos futuros.

CAPÍTULO

2

CONCEITOS RELACIONADOS

Neste capítulo, são apresentados os conceitos fundamentais para o melhor entendimento da proposta deste trabalho. Nesse sentido, a Seção 2.1 apresenta a definição de vídeo digital e sua estrutura temporal hierárquica em quadros, tomadas e cenas.

O processo de segmentação em cenas, englobando as etapas geralmente adotadas por uma técnica de segmentação em cenas, é detalhado na Seção 2.2.

Já a Seção 2.3 apresenta o processo de extração de características, adotado para a redução do grande volume de dados presente em um vídeo, descrevendo ainda três características de diferentes modalidades amplamente utilizadas por técnicas reportadas na literatura.

A Seção 2.4 apresenta a Aprendizagem Profunda, uma subárea da Aprendizagem de Máquina que vem sendo empregada com sucesso em tarefas relacionadas, como classificação de imagens e detecção de eventos em vídeos. Neste trabalho é esperado que a mesma possa contribuir auxiliando a encontrar correlações entre tomadas adjacentes, auxiliando na identificação de tomadas pertencentes a uma mesma cena.

Por se tratar de uma abordagem baseada em informações de diferentes modalidades, um processo de fusão multimodal foi adotado neste trabalho. É esperado que informações individuais em cada modalidade possam ser extraídas e representadas de maneira complementar, semanticamente mais ricas. Além disso, espera-se também que tal representação favoreça a identificação de correlações temporais entre tomadas promovendo aumento de eficácia na tarefa de segmentação em cenas. A descrição do processo de fusão multimodal, assim como as duas abordagens geralmente adotadas pelos pesquisadores da área, é apresentada na Seção 2.5.

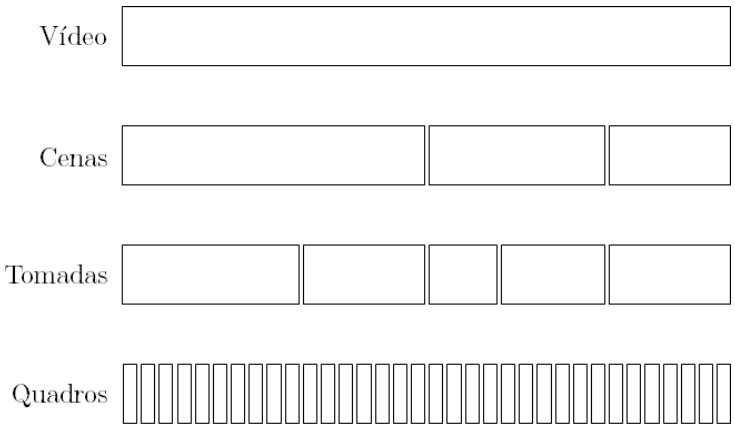
Por sua vez, são apresentadas e discutidas na Seção 2.6 considerações acerca da avaliação de eficácia de técnicas de segmentação em cenas, especificamente as métricas adotadas e bases de dados disponíveis para utilização. Por fim, a Seção 2.7 apresenta algumas considerações sobre os conceitos discutidos ao longo deste capítulo.

2.1 Video Digital

Um vídeo digital, segundo Blanken *et al.* (2007), pode ser definido como uma sequência de imagens ou quadros apresentados a uma determinada taxa, transmitindo a sensação de movimento. Cada quadro do vídeo é representado como uma matriz, no qual o número de linhas forma a resolução vertical e o número de colunas a resolução horizontal, determinando o número de pixels do quadro. Cada pixel de um quadro é geralmente representado por três bytes, usando o espaço-de-cor RGB (*Red-Green-Blue*), no qual cada byte indica a intensidade, entre 0 e 255, que o pixel possui de cada componente de cor (R, G e B). Vídeos compactados podem ainda ser armazenados seguindo algum esquema adotado com o intuito de reduzir o volume de dados, como a subamostragem de crominância em espaços de cor YCbCr, entre outros.

Semanticamente, um vídeo digital está estruturado hierarquicamente, conforme ilustrado na Figura 2, em quadros, tomadas, cenas e o vídeo em si.

Figura 2 – Estrutura hierárquica de um vídeo digital em quadros, tomadas e cenas.



Fonte: Adaptada de Coimbra e Goularte (2009).

Os quadros podem ser definidos como uma única imagem estática (RICHARDSON, 2002) que, quando apresentados a uma determinada taxa dão a ilusão do movimento. Com o desenvolvimento de codificadores estabelecidos, a segmentação em quadros é considerada trivial, com a presença de ferramentas robustas em um amplo número de situações.

Já as tomadas podem ser definidas como um conjunto de quadros estáticos obtidos continuamente por uma única câmera (KOPRINSKA; CARRATO, 2001). Sua segmentação, considerada resolvida (FABRO; BÖSZÖRMENYI, 2013), é consideravelmente mais complexa que a segmentação em quadros, se baseando na análise de informações de baixo nível para identificar descontinuidades entre quadros adjacentes. Exemplos de informações que podem sugerir a transição entre tomadas são uma mudança repentina no ângulo da câmera, desaparecimento de determinados objetos sendo capturados ou até mesmo modificações significativas na cor dos objetos representados.

Já as cenas, também chamadas de *logical story unit* (HANJALIC; LAGENDIJK; BIE-

MOND, 1999; SIDIROPOULOS *et al.*, 2011), *story unit* (BOLLE; YEO; YEUNG, 1998), ou simplesmente *story* (BARALDI; GRANA; CUCCHIARA, 2017), não possuem uma definição consensual, sendo usualmente definidas como um conjunto de tomadas adjacentes semanticamente relacionadas (LIANG *et al.*, 2009; BABER; AFZULPURKAR; BAKHTYAR, 2011; CHASANIS; LIKAS; GALATSANOS, 2009; WANG *et al.*, 2006). Essa também é definição adotada neste trabalho. Segundo Fonseca (2006), o agrupamento das tomadas em cenas depende do julgamento subjetivo da correlação semântica, o que pode exigir a análise de informações além das visuais.

É importante ressalvar que, embora amplamente adotada, a segmentação hierárquica previamente mencionada não é a única reportada na literatura. Dependendo do domínio ou objetivo do vídeo, segmentações diversas podem ser efetuadas, como segmentação por notícias em vídeos de telejornais, segmentação em tópicos em video-aulas, segmentação em capítulos em filmes de longa duração e até mesmo a segmentação em trechos ou partes específicas de casamentos (SAWAI *et al.*, 2011), entre outras.

Devido ao alto volume de informações presente em um vídeo e sua particularidade de ser um tipo de informação não estruturada, os pesquisadores utilizam um processo de extração de características (do inglês *feature extraction*) dos dados presentes no vídeo, conforme descrito na Seção 2.3.

2.2 Processo de segmentação em cenas

O processo de segmentação em cenas, conforme geralmente adotado por técnicas reportadas na literatura, é geralmente composto de três etapas diferentes e sequenciais: Pré-processamento, Segmentação e Pós-processamento (Figura 1).

A etapa de Pré-processamento se refere ao processo de extração de informações relevantes do vídeo que serão posteriormente utilizadas para identificar as cenas. Tal etapa engloba ainda a seleção ou filtragem de informações a serem utilizadas, processamento relativo a redução da dimensionalidade ou volume de dados, normalização das informações entre diferentes modalidades, a construção de uma representação adequada para cada modalidade de entrada, entre outros.

A segunda etapa, Segmentação, se refere a todos os processos realizados com o objetivo de identificar ou agrupar as tomadas que pertencem a uma mesma cena. Nesse sentido, uma ampla gama de técnicas são reportadas na literatura, incluindo técnicas de reconhecimento de padrões (CHASANIS; LIKAS; GALATSANOS, 2009), identificação de correlações em grafos conectados (SIDIROPOULOS *et al.*, 2011), classificação e agrupamento de tomadas (BARALDI; GRANA; CUCCHIARA, 2015a), cálculo da similaridade de informações visuais (RASHEED; SHAH, 2003), entre outros. É importante destacar ainda que tal etapa é altamente atrelada a anterior, sendo comum o uso de uma técnica específica de segmentação em cenas devido

exclusivamente a particularidades das informações ou características de entrada.

A terceira e última etapa, Pós-processamento, consiste em um conjunto de técnicas que podem ser aplicadas com o intuito de detectar e remover possíveis equívocos na segmentação em cenas obtida na etapa anterior, potencialmente obtendo ganhos significativos em termos de eficácia. Rasheed e Shah (2003), por exemplo, removem a ocorrência de falsos positivos da segmentação em cenas de ação por meio da análise da *quantidade de movimento* presente em tomadas de baixa duração.

Assim, atualmente, um pesquisador que queira criar um método para obter a segmentação automática em cenas deve escolher um conjunto de informações ou características a serem extraídas dos vídeos de entrada, gerando uma representação adequada que será utilizada posteriormente. A seguir, é criado um método de segmentação relevante e condizente com a representação das informações extraídas anteriormente, obtendo uma segmentação em cenas preliminar. Por fim, uma série de processos podem ser adicionados a segmentação candidata de maneira a remover ruídos ou comportamentos que prejudiquem a qualidade da segmentação.

2.3 Extração de características

Segundo Duda, Hart e Stork (2000), o processo de extração de características é um processo aplicado para reduzir o grande volume de dados que um vídeo possui ao eliminar informações possivelmente redundantes. Tais características podem, segundo Blanken *et al.* (2007), ser classificadas como de alto ou baixo nível semântico.

Características de baixo nível semântico podem ser obtidas automaticamente pela análise de padrões ou estatística de um objeto multimídia, sendo, portanto, altamente dependentes do tipo de dados em análise. Usando os quadros do vídeo, por exemplo, é possível calcular a quantidade de pixels que apresentam cores similares, dando origem aos histogramas de cor (GONZALEZ; WOODS, 2009). Por sua vez, usando as informações do áudio do vídeo, formado pela amplitude amostrada digitalmente da pressão do ar, é possível detectar informações de baixo nível como intensidade média ou a mudança do sinal no chamado *Zero Crossing Rate* (ZCR) (GOUYON; PACHET; DELERUE, 2000).

Características de alto nível, por outro lado, possuem informações relevantes e compreensíveis pelos usuários. Exemplos de tais características são as *tags* ou rótulos que identificam o significado ou a presença de objetos, assuntos ou ações, que podem ser usadas para auxiliar em tarefas como indexação ou recomendação de conteúdo. Em vídeos digitais, tais informações podem estar disponíveis de acordo com o meio de transmissão ou armazenamento do vídeo, incluindo marcadores de capítulos em mídias físicas (como BluRays) ou em serviços online de visualização de vídeos como YouTube ou NetFlix. Tais características, porém, são adicionais ao vídeo em si, podendo estar ausente de acordo com a mídia utilizada.

Claramente, há uma grande distância entre a representação computacional dos dados e o seu significado, sendo tal problema conhecido como lacuna semântica (SMEULDERS *et al.*, 2000).

As características de baixo nível normalmente podem ser extraídas automaticamente e são usadas amplamente na tarefa de segmentação. São identificadas por uma representação, chamada de descritores de características, que define sua sintaxe e semântica (WU *et al.*, 1999). Um descritor é definido como uma tupla (e, D), no qual e é um extrator de características e D uma medida de dissimilaridade. Um extrator de características é um algoritmo ou função que recebe um determinado conjunto de dados, como os valores de pixels de uma imagem, retornando uma representação compacta de seu conteúdo, como um histograma de cor da imagem. Essa representação, chamada de vetor de característica, é, portanto, um vetor numérico, geralmente de dimensionalidade preestabelecida. Por sua vez, a medida de dissimilaridade tem por objetivo medir a dissimilaridade entre vetores de características distintos.

A seguir são aprofundandos os três métodos de extração de características visual, aural e textual em destaque na literatura, sendo amplamente utilizados em trabalhos relacionados por apresentarem bons resultados em tarefas diversas.

2.3.1 Scale Invariant Feature Transform (SIFT)

Proposto por Lowe (2004), o *Scale Invariant Feature Transform* (SIFT) é um extrator de características locais de imagens empregado em diversas tarefas que envolvem processamento de imagens, inclusive na segmentação em cenas (CHASANIS; KALOGERATOS; LIKAS, 2009; LOPES; TROJAHN; GOULARTE, 2014). Ao analisar a informação de sua vizinhança, o SIFT é capaz de encontrar pontos específicos que sejam altamente descritivos de uma imagem, sendo tais pontos relativamente resistentes a transformações geométricas e de escala.

Dada uma imagem em escala-de-cinza de entrada, o algoritmo realiza quatro operações: detecção de extremos, localização de pontos-chave (do inglês *key points*), definição de orientação e descrição dos pontos-chave. Os dois primeiros consistem na parte relativa ao detector de características, enquanto que os dois últimos consistem no descritor SIFT em si.

A detecção de extremos visa identificar pontos na imagem que sejam resistentes a mudanças de escala. Para isso, é utilizada uma função Gaussiana definida como:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \cdot e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{2.1}$$

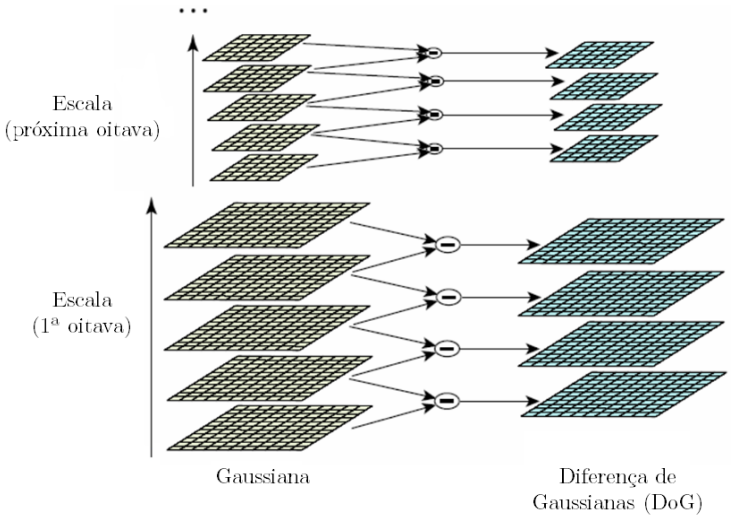
Onde x e y são os índices da imagem e σ é um parâmetro que define a escala do filtro.

Define-se, então, a Diferença de Gaussianas (do inglês *Difference of Gaussian* - DoG) como a diferença entre dois filtros Gaussianos separados por um valor de escala k tal que:

$$DoG = G(x, y, k\sigma) - G(x, y, \sigma) \tag{2.2}$$

O filtro Gaussiano é aplicado na imagem de entrada em diferentes escalas σ para remover ruídos e detalhes indesejáveis, mantendo apenas possíveis pontos de interesse. Variando o valor da escala σ, são obtidas a DoG em escalas diferentes, gerando uma oitava (do inglês *octave*). Para o cálculo da próxima oitava, é utilizada a imagem em escala central após aplicada uma subamostragem para reduzir sua resolução. Lowe (2004) definem empiricamente o uso de três escalas σ em cada oitava, valor de σ igual a 1.6 e um total de quatro oitavas (dependentes da resolução da imagem de entrada). A Figura 3 ilustra o cálculo das Gaussianas e da DoG formando as oitavas em cada escala.

Figura 3 – Ilustração do cálculo das Gaussianas e da Diferença de Gaussianas (DoG) em diferentes escalas formando as oitavas no extrator de características SIFT



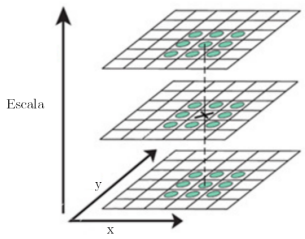
Fonte: Adaptada de Lowe (2004).

Após extraídos as DoG em cada oitava, são selecionados os pontos extremos que apresentem o maior ou menor valor que todos seus vizinhos tanto espacialmente como nas escalas adjacentes, conforme ilustrado na Figura 4. Tais pontos extremos ou pontos-chaves candidatos são localizados na imagem por meio da aplicação da expansão de Taylor da função DoG (BROWN; LOWE, 2002). Por fim, os pontos-chaves são analisados por meio de uma função de estabilidade e descartados caso sejam considerados instáveis.

Para gerar descritores invariantes quanto a rotação, são atribuídas orientações para cada ponto-chave localizado. Usando a diferença de pixels, calcula-se um histograma de orientação de 36 posições (de 0 a 2π) cujo valor de peso é dado pela magnitude do gradiente. Os valores de pico no histograma indicam orientações dominantes, sendo considerados todos os valores acima de pelo menos 80% do valor máximo encontrado. A Figura 5 ilustra a construção do histograma de orientações dado um ponto-chave detectado.

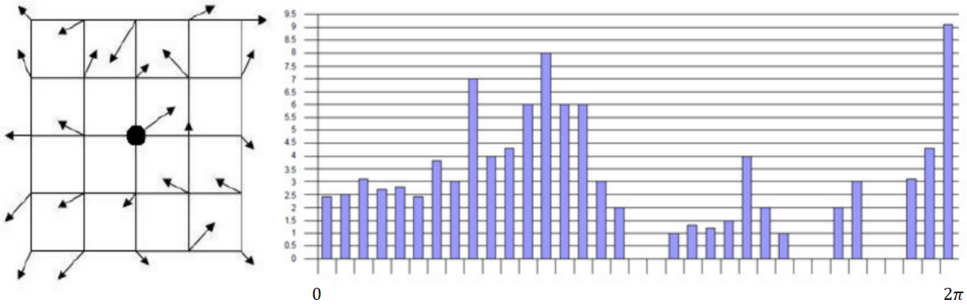
É importante destacar que um ponto-chave pode possuir mais de uma orientação, sendo o valor máximo (ou a interpolação da parábola entre os três maiores valores) utilizado como orientação do ponto-chave em si.

Figura 4 – Ilustração da seleção de um ponto-chave dada sua vizinha espacial em diferentes escalas no extrator de características SIFT



Fonte: Adaptada de Lowe (2004).

Figura 5 – Ilustração do histograma de orientações de um ponto-chave no extrator de características SIFT



Fonte: Adaptada de Lowe (2004).

Já na construção do descritor, uma região de tamanho $16x16$ é definida ao redor do ponto-chave. A orientação de cada pixel é suavizada utilizando uma função Gaussiana com escala igual a metade do tamanho da janela, reduzindo o impacto de pequenas oscilações da orientação e dando prioridade a orientações próximas ao centro do descritor, menos propensas a ruído. Após a suavização, um histograma normalizado de 8 posições é gerado para cada grupo de $4x4$ pixels, totalizando 16 histogramas. Por fim, o descritor final para cada ponto-chave é formado pela concatenação de tais histogramas, resultando em um vetor único de 128 posições.

2.3.2 Mel-Frequency Cepstrum Coefficients (MFCC)

Dentre os diferentes extratores de características aurais reportados na literatura como o *Linear Predictor Coefficient* (LPC), *Linear Predictive Cepstral Coefficients* (LPCC) e *Linear Frequency Cepstral Coefficient* (LFCC) (ATREY; MADDAGE; KANKANHALLI, 2006), o *Mel-Frequency Cepstrum Coefficients* (MFCC) é de especial interesse devido a sua popularidade em diversas tarefas de áudio (MÜLLER, 2007; SAHIDULLAH; SAHA, 2012). Seu funcionamento é baseado no sistema auditivo humano, no qual a frequência é posicionada logaritmicamente na chamada escala *mel* (STEVENS; VOLKMANN; NEWMAN, 1937), sendo especialmente eficiente na captura do timbre do áudio (MÜLLER, 2007).

As características MFCC podem ser obtidas de um sinal de áudio, segundo Hong (2017),

por meio dos seguintes passos:

1. Os dados de entrada são agrupados em janelas, geralmente com tamanhos entre 20ms a 40ms (HONG, 2017), de modo a obter amostras estáveis em cada janela. Um valor de janela excessivamente baixo pode levar a ausência de amostras significativas em cada janela, enquanto que valores excessivamente altos podem resultar em demasiada variação entre cada janela. Cada janela é calculada tendo uma certa sobreposição entre janelas adjacentes.

2. Após cada janela de amostras, é aplicado um processo de janelamento de Hamming (do inglês *Hamming window*) de modo a minimizar possíveis descontinuidades antes e após cada janela.

3. A Transformada Discreta de Fourier (do inglês *Discrete Fourier Transform* - DFT) que converte cada janela do domínio temporal para o domínio de frequências, é cálculada. O resultado, chamado de espectro de frequência do sinal, é geralmente obtido pelo algoritmo de Transformada Rápida de Fourier (do inglês *Fast Fourier Transform* - FFT).

4. A seguir, um banco de filtros de Mel são aplicados, medindo a energia do sinal em diferentes frequências. Como o ouvido humano é incapaz de perceber linearmente o aumento da energia do sinal, o logaritmo do resultado de cada filtro é calculado.

5. Por fim, a DCT do logaritmo de cada filtro aplicado anteriormente é claculada. Dos coeficientes resultantes são selecionados os treze primeiros (incluindo o elemento DC e os próximos doze coeficientes), no qual os demais são descartados para evitar a influência de sinais pouco significativos. Assim, o descritor MFCC é composto de um vetor de 13 dimensões para cada janela calculada.

A literatura não aponta nenhuma medida de dissimilaridade consensual para a comparação entre descritores MFCC. Nesse sentido, alguns autores utilizam a distância euclidiana, tal como Majeed *et al.* (2015), ou ainda o *Dynamic Time Warping* (DTW) para analisar trechos com diferentes durações. O MFCC é considerado eficiente e resistente a diversos tipos de ruídos, sendo amplamente utilizado para sistemas de reconhecimento de voz (SAHIDULLAH; SAHA, 2012) como, por exemplo, a detecção da identidade da voz (GANCHEV; FAKOTAKIS; KOKKINAKIS, 2005) e outras tarefas baseadas na modelagem da voz humana, inclusive na segmentação de vídeo em cenas (WU; JIN, 2015; SIDIROPOULOS *et al.*, 2011; LOPES; TROJAHN; GOULARTE, 2014; BARALDI; GRANA; CUCCHIARA, 2015a; BARALDI; GRANA; CUCCHIARA, 2017).

2.3.3 Bag of Words (BoW)

Dentre diversas abordagens reportadas na literatura para a análise textual, uma abordagem popular é o modelo *Bag of Words* (BoW). Desenvolvido para a análise de *corpus* textuais na

área de processamento de linguagem natural, o BoW é um modelo que consiste em obter uma representação de um documento $D = (t_1, t_2, ..., t_p)$ no qual p é o tamanho do vocabulário e t_i é a relevância do termo t_i no documento D (SALTON; BUCKLEY, 1988). Cada termo no modelo BoW pode se referir a apenas uma única palavra, a duas palavras (bigrama) (TAN; WANG; LEE, 2002) ou até mesmo uma frase completa (LEWIS, 1992). Já a relevância do termo pode ser calculada de diversas formas, incluindo a frequência do termo (do inglês *term frequency* - tf) ou a frequência inversa do termo (do inglês *inverse term frequency* - idf) (KO, 2012; KO, 2015).

Considerando $f_{t,d}$ como a frequência do termo t no documento d, a frequência do termo (tf) pode ser obtida de diversas formas (MANNING; RAGHAVAN; SCHüTZE, 2008), como a definida a seguir:

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{2.3}$$

Ao invés de contar o número de ocorrências de cada termo no documento, a frequência inversa do termo (idf) proposta por JONES (1972) busca medir o número de documentos de um *corpus* em que o termo ocorre, separando palavras comuns e que aparecem na maioria dos documentos ou palavras raras que aparecem em poucos documentos. Considerando D o conjunto de documentos do corpus textual e $|D|$ o número de documentos em D e $|d \in D : t \in d|$ o número de documentos que possuem o termo t, a frequência inversa do termo $idf(t,D)$ pode ser calculada como:

$$idf(t,D) = \log \frac{|D|}{1 + |d \in D : t \in d|} \tag{2.4}$$

Além do cálculo da frequência de um termo, um importante aspecto na aquisição do conhecimento por informações textuais não-estruturadas é a etapa de pré-processamento (REZENDE; MARCACINI; MOURA, 2011), no qual os termos de entrada são tratados e padronizados, preservando as principais características do texto de entrada. Exemplos de abordagens de pré-processamento amplamente utilizadas são a remoção de palavras muito comuns, chamadas de *stop words*, padronização do texto para minúsculas, radiciação (do inglês *stemming*) e lematização (do inglês *lemmatization*), entre outros. Nesse sentido, Uysal e Gunal (2014) reporta uma avaliação de diferentes técnicas de pré-processamento e seu impacto em diversos *corpus* textuais diferentes.

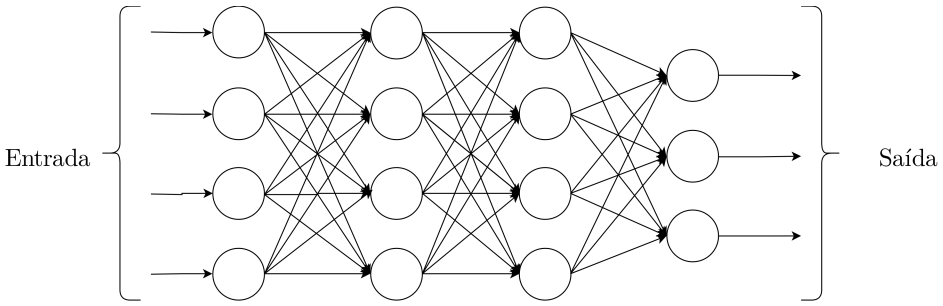
No caso particular da segmentação em cenas, busca-se representar o conteúdo textual de um dado segmento de vídeo (por exemplo: tomada) por um vetor de característica. Baraldi, Grana e Cucchiara (2015a), por exemplo, utilizam redes Word2Vec (Mikolov *et al.*, 2013) para gerar um vetor de característica para cada termo da legenda de uma tomada. Após convertidos, os vetores de características que representam os termos extraídos são então agrupados, formando um dicionário textual, que é utilizado para gerar um histograma normalizado de frequência das palavras contidas em cada legenda. Para evitar casos no qual legendas com duração de múltiplas tomadas sejam representadas em uma única tomada, que resultaria em vetores de características bastante esparsos, Baraldi, Grana e Cucchiara (2015a) utilizaram uma janela

de tolerância (chamada de janela de contexto) de 20 segundos de duração. Tanto o número de palavras selecionadas para o dicionário textual como o número de dimensões do histograma de frequência dos termos de cada legenda possuem tamanho 200.

2.4 Aprendizagem Profunda

A Aprendizagem Profunda, termo inicialmente adotado por Dechter (1986), se refere a um conjunto de técnicas relacionadas a redes neurais com múltiplas camadas. Uma rede neural, segundo Gurney (1997) é um conjunto interconectado de elementos simples de processamento, chamados de unidades, sendo sua habilidade de processamento armazenada nas interconexões entre as mesmas, chamadas de pesos (do inglês *weights*), cujos valores são obtidos por um processo de adaptação ou aprendizagem dado um conjunto de padrões ou exemplos de treinamento. As unidades das rede neurais são organizadas em camadas, nas quais há uma conexão ou sinapse entre as unidades de camadas adjacentes, mas não entre unidades de uma mesma camada (MARTINEZ, 1996). Nesse sentido, a Figura 6 ilustra uma rede neural de quatro camadas.

Figura 6 – Ilustração de uma rede neural organizada em quatro camadas, na qual cada círculo representa uma unidade ou neurônio e cada seta uma conexão ou sinapse contendo seu respectivo peso



Fonte: Adaptada de Haykin (2009).

A popularidade recente da Aprendizagem Profunda é fruto de diversos fatores, como o desenvolvimento de abordagens bem-sucedidas para diferentes tarefas, da criação de *frameworks* de desenvolvimento de alta eficiência e de fácil utilização, do aperfeiçoamento e utilização de *hardware* dedicado por meio de bibliotecas de alto desempenho e da proliferação de grandes bases de dados adequadas ao treinamento de redes neurais. Tais avanços permitiram o desenvolvimento e treinamento de redes com maior poder de aprendizagem e representatividade, obtendo considerável sucesso em diversas tarefas, como a classificação de imagens (RUSSAKOVSKY *et al.*, 2015; HE *et al.*, 2015) e vídeos (WU *et al.*, 2016), reconhecimento de fala (GRAVES; JAITLY, 2014), entre outros.

Duas abordagens da Aprendizagem Profunda são de especial interesse para a proposta deste trabalho: as redes neurais convolucionais (CNN) e as redes neurais recorrentes (RNN). As redes convolucionais são amplamente adotadas para a extração e aprendizagem de padrões de alto nível dos dados de entrada. Já as redes recorrentes são aplicáveis para a análise de dados

sequenciais de tamanho variável. As redes convolucionais e recorrentes são descritas nas Seções subsequentes, respectivamente.

2.4.1 Redes convolucionais

As redes neurais convolucionais são redes neurais de múltiplas camadas, compostas de camadas que realizam as operações de convolução e subamostragem (PRASOON *et al.*, 2013), geralmente adotadas para a extração de padrões distintos (WIATOWSKI; BOLCSKEI, 2018). A *convolução*, quando aplicada ao processamento digital de imagens, pode ser entendida como o processo de mover uma máscara ou *kernel* pela imagem e calcular a soma dos produtos em cada posição (GONZALEZ; WOODS, 2009).

Atualmente, as CNN, segundo Wiatowski e Bolcskei (2018), são formadas por combinações de camadas convolucionais, camadas não-lineares e camadas de *pooling* ou subamostragem, responsáveis por prover melhor adaptabilidade e redução de dimensionalidade. Seu resultado, um conjunto de padrões obtidos dos dados de entrada, podem então ser utilizados por uma rede neural para fins diversos, como classificação de imagens, no qual obtém considerável sucesso (SIMONYAN; ZISSERMAN, 2014).

Formalmente, a saída x_{out} de uma camada convolucional sobre um valor de entrada x_{in} sob uma máscara ou *kernel* k, uma *bias* escalar opcional b e uma função não-linear S, é dado por (HUANG; LECUN, 2006):

$$x_{out} = S\left(\sum_i x_{in} \circledast k_i + b\right) \tag{2.5}$$

Onde $\circledast$ é a operação de convolução. É importante notar que a Equação 2.5, que define a convolução sobre dados unidimensionais 1D, pode ser estendida para dados de maior dimensionalidade. Nesse caso, considerando uma matriz 2D x_{in} e uma máscara ou *kernel* 2D k centrada na posição de índice $[0,0]$, a operação de convolução $x_{in}[m,n] \circledast k$ é definida como:

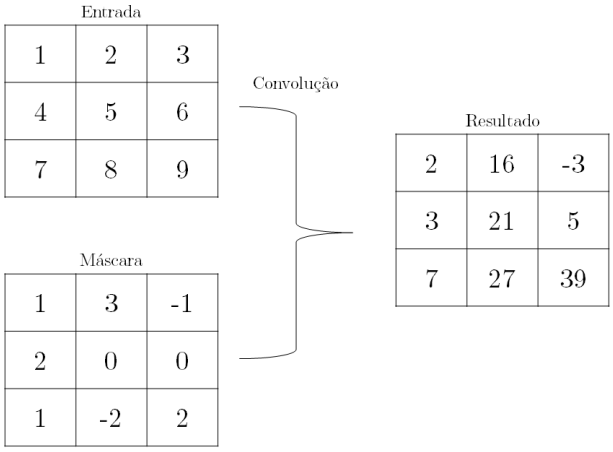
$$x_{in}[m,n] \circledast k = \sum_{j=-\infty}^{\infty} \sum_{i=-\infty}^{\infty} x_{in}[i,j] \cdot k[m-i,n-i] \tag{2.6}$$

Como exemplo, a Figura 7 ilustra a aplicação de uma convolução 2D ($x_{in} \circledast k_i$) dada uma matriz 3x3 de entrada com uma máscara predefinida de mesmo tamanho.

Assim, a camada convolucional tem por objetivo estimar e aplicar uma máscara sobre os dados de entrada, realizando operações tais como a detecção de bordas, filtros passa-baixa ou passa-alta, entre outras. É importante notar ainda que cada camada convolucional pode aplicar uma combinação de diversas máscaras diferentes (HUANG; LECUN, 2006).

Segundo Sachdeva *et al.* (2017), é necessário a utilização de funções não-lineares após a operação de convolução para evitar que sucessivas camadas convolucionais operem como uma simples operação linear, resultando em baixa capacidade de aprendizagem e uma eficácia inadequada.

Figura 7 – Exemplo da aplicação da operação de convolução 2D sobre uma matriz de entrada de tamanho 3x3 com uma máscara predefinida de mesmo tamanho.



Fonte: Elaborada pelo autor.

Uma dificuldade que redes neurais enfrentam é a sobre-especialização (do inglês *overfitting*), no qual uma rede neural treinada se torna incapaz de classificar adequadamente amostras inéditas, apresentando uma baixa capacidade de generalização. Segundo Hochreiter (1998), um dos fatores que levam a tal problema é o uso de funções não-lineares como a sigmoide, cujo valor de saída tende a desaparecer (valor se aproxima de zero) com a execução sucessiva em múltiplas camadas da rede. Para resolver tal limitação pesquisadores desenvolveram diferentes funções retificadoras como a *Rectified Linear Unit* (ReLU) (NAIR; HINTON, 2010), amplamente utilizada em trabalhos recentes que utilizam redes convolucionais. Dado um valor de entrada a na forma de $a = W \cdot x + b$, no qual W é o peso da conexão ou sinapse do valor de entrada x e b um valor de *bias* opcional, a ReLU pode ser definida como:

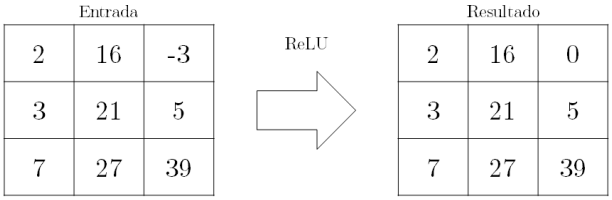
$$ReLU = \max(0, a) \tag{2.7}$$

Segundo Krizhevsky, Sutskever e Hinton (2012), a popularidade da função ReLU advém de seu alto desempenho e sua maior resistência a sobre-especialização quando comparado a outras funções não-lineares como a sigmoide. Além dela, pesquisadores desenvolveram ainda uma série de novas funções retificadoras como a *Leaky Rectified Linear Unit* (LReLU) (MAAS; HANNUN; NG, 2013), *Parametric Rectified Linear Unit* (PReLU) (HE *et al.*, 2015), entre outros (XU *et al.*, 2015). Tais funções tem como ideia básica manter uma maior esparcidade quando comparada com funções não-lineares tradicionais, tendo como característica comum retornar o valor de entrada caso o mesmo seja igual ou maior que zero, adotando diferentes abordagens caso contrário. Por exemplo, dado a na forma de $a = W \cdot x + b$ de maneira idêntica a definição da ReLU (Equação 2.7) e α, um parâmetro predefinido, a LReLU pode ser definida como:

$$LReLU = \begin{cases} \alpha \cdot a, & \text{se } a < 0 \\ a, & \text{caso contrário} \end{cases} \tag{2.8}$$

A Figura 8 ilustra a execução da ReLU sobre o resultado da convolução ilustrada na Figura 7, considerando um valor de *bias* zero e o valor unitário para os pesos de cada conexão entre a camada convolucional e a camada não-linear (ou seja, $a = 1 \cdot x + 0$ para cada valor x resultante da operação de convolução).

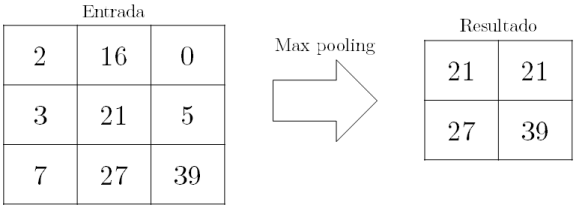
Figura 8 – Exemplo da aplicação da função retificadora ReLU sobre uma matriz 3x3 de entrada.



Fonte: Elaborada pelo autor.

Finalmente, após a execução da operação de convolução e da função não-linear, representadas por x_{out} na Equação 2.5, geralmente é aplicada uma operação de subamostragem ou *pooling* visando a reduzir a dimensionalidade, reduzindo o número de parâmetros ou valores de entrada (SACHDEVA *et al.*, 2017), além de aumentar a invariância à distorção (HUANG; LECUN, 2006) e à translação (GOODFELLOW; BENGIO; COURVILLE, 2016) quando aplicadas a imagens. Dentre as diversas operações de subamostragem existentes, a chamada *max pooling* é especialmente popular (SIMONYAN; ZISSERMAN, 2014; HE *et al.*, 2016). A Figura 9 ilustra a execução da operação de *max pooling* em uma janela 2x2 sobre a matriz 3x3 obtida como resultado na Figura 8, considerando um deslocamento de janelas deslizantes de tamanho 1.

Figura 9 – Exemplo da aplicação da função de subamostragem *max pooling* em uma janela 2x2 sobre uma matriz 3x3 de entrada.



Fonte: Elaborada pelo autor.

A capacidade de aprender padrões significativos sobre dados brutos de entrada (WU *et al.*, 2016), sem a necessidade de interferência humana, torna as CNNs uma das abordagens mais populares da Aprendizagem Profunda. Tais redes foram empregadas neste trabalho para o reconhecimento de padrões entre as características extraídas de cada modalidade, descritas na Seção 2.3, com o objetivo de criar um vetor de característica único para cada tomada do vídeo de entrada.

2.4.2 Redes neurais recorrentes

As redes neurais recorrentes (RUMELHART; HINTON; WILLIAMS, 1988; ELMAN, 1990) se diferenciam das redes neurais tradicionais por possuir ao menos uma camada dita recorrente, cuja saída é utilizada também como entrada para a mesma camada. Ou seja, se trata de uma camada cujo valor de saída ou ativação é resultado, além do dado de entrada, da saída da própria unidade na iteração (tempo) anterior. Tal característica torna a rede capaz de processar informações sequenciais (GOODFELLOW; BENGIO; COURVILLE, 2016), especialmente temporais, potencialmente de tamanho variável (CHUNG *et al.*, 2014).

Especificamente, seja x_t o valor de entrada no tempo t, $\sigma(h_{t-1})$ o resultado da rede no tempo $t-1$, W_* o vetor com os pesos das conexões ou sinapses de entrada e b um valor de *bias* opcional, então o estado da rede h no tempo t é dado por (PASCANU; MIKOLOV; BENGIO, 2013):

$$h_t = W_{rec}\sigma\left(h_{t-1}\right) + W_{in}x_t + b \tag{2.9}$$

A atualização dos pesos ou parâmetros de redes neurais recorrentes é feito, geralmente, por meio do algoritmo *Backpropagation Through Time* (BPTT) (BEAUFAYS; ABDEL-MAGID; WIDROW, 1994; NGUYEN; WIDROW, 1990; WERBOS, 1990). Similar ao algoritmo *backpropagation*, o BPTT se baseia em desdobrar a rede recorrente em relação ao tempo, calculando o gradiente dos pesos ou parâmetros da rede a cada passo ou iteração.

As redes neurais recorrentes são amplamente utilizadas em diversas tarefas, tal como a detecção de bordas de estórias em documentos textuais (YU *et al.*, 2016; YU *et al.*, 2017; TSUNOO; BELL; RENALS, 2017; TSUNOO *et al.*, 2017), predição de pontuação (XU; XIE; YAO, 2016), modelagem de fala (TAN *et al.*, 2016), entre outros. Li e Wu (2015), na tarefa de reconhecimento de fala, conclui que uma abordagem baseada em RNNs obtém resultados em estado da arte superiores a uma abordagem de Aprendizagem Profunda não-recorrente. Segundo Yu *et al.* (2017), a principal vantagem de uma RNN é a facilidade de se modelar o contexto, enquanto que redes tradicionais requerem a adição de informações de contexto da entrada por meio de janelas deslizantes.

Duas limitações importantes das RNNs são amplamente mencionadas por pesquisadores: a explosão do gradiente (do inglês *exploding gradient*) (BENGIO; SIMARD; FRASCONI, 1994) e o desaparecimento do gradiente (do inglês *vanishing gradient*) (JOZEFOWICZ; ZAREMBA; SUTSKEVER, 2015). Em ambos os casos, ao aplicar o BPTT, o gradiente aplicado a atualização dos pesos da rede tende a aumentar (explosão) ou diminuir (desaparecer) rapidamente conforme o aumento no número de iterações do algoritmo. Isso torna uma RNN incapaz de tratar adequadamente dados em sequências longas, o que prejudica sua eficácia em diversas tarefas. Uma análise detalhada de tais problemas é descrita no trabalho de Pascanu, Mikolov e Bengio (2013).

Um método para tratar da explosão do gradiente, atualmente, é reduzir ou remover os gradientes que excederem determinado limiar (PASCANU; MIKOLOV; BENGIO, 2013). Já o

desaparecimento do gradiente, conhecido por ser um problema mais desafiador para as RNNs convencionais (JOZEFOWICZ; ZAREMBA; SUTSKEVER, 2015), pode ser evitado por meio do uso de uma arquitetura com um padrão de conectividade específico. Exemplos de arquiteturas que adotam tal estratégia são a *Gated Recurrent Unit* (GRU) e *Long Short Term Memory* (LSTM), desenvolvida por Hochreiter e Schmidhuber (1997) e considerada a abordagem padrão para redes recorrentes (JOZEFOWICZ; ZAREMBA; SUTSKEVER, 2015).

Na abordagem proposta por Hochreiter e Schmidhuber (1997) cada unidade de processamento da rede é composta por diversos *gates*, conectados de maneira específica, formando uma célula de memória (do inglês *memory cell*) que armazena informações sobre as últimas n execuções. É importante destacar que diferentes autores empregam pequenas alterações nos diferentes elementos da LSTM. Neste trabalho foi utilizada a formulação da LSTM definida por Jozefowicz, Zaremba e Sutskever (2015), baseada no trabalho de Graves (2013), que omite algumas interconexões específicas entre *gates*, obtendo ganhos de desempenho e mantendo um equivalente poder de representatividade.

Seja W_* o vetor com os pesos das conexões ou sinapses, b_* a sua respectiva *bias*, h_t o valor de saída de um elemento LSTM no tempo t, x_t o valor de entrada para o elemento LSTM no tempo t, $sigm$ a função sigmoide e $tanh$ a função tangente hiperbólica, assim, a função *forget gate* f_t, responsável por determinar o quanto a entrada deverá ser memorizada, é definida como:

$$f_t = sigm\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \tag{2.10}$$

Analogamente, os *gates* de entrada i_t e j_t, usados para armazenar o estado da célula de memória c_t no tempo t, são definidos como:

$$i_t = tanh\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right) \tag{2.11}$$

$$j_t = sigm\left(W_{xj}x_t + W_{hj}h_{t-1} + b_j\right) \tag{2.12}$$

Assim, usando os *gates* previamente definidos, é possível atualizar o estado da célula c_{t-1} para c_t definido como:

$$c_t = c_{t-1} \odot f_t + i_t \odot j_t \tag{2.13}$$

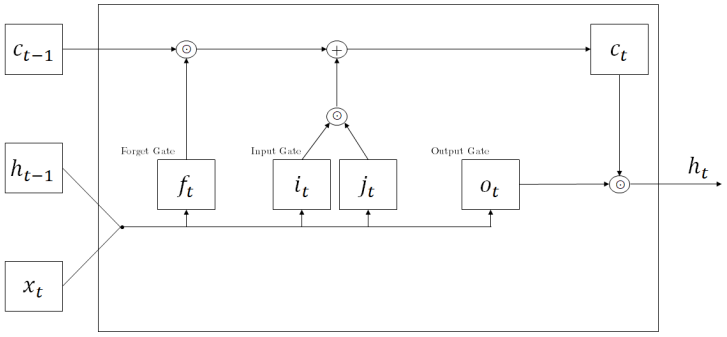
Onde $\odot$ é o produto vetorial de matrizes ponto-a-ponto. Por fim, a saída h_t do elemento LSTM no tempo t, calculado por meio do *gate* de saída o_t é definido como:

$$o_t = tanh\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right) \tag{2.14}$$

$$h_t = tanh\left(c_t\right) \odot o_t \tag{2.15}$$

A Figura 10 ilustra o padrão de conectividade previamente mencionado, com destaque no *memory cell*.

Figura 10 – Ilustração de uma unidade LSTM, no qual $\odot$ é o produto vetorial entre matrizes, x_t é o valor de entrada e h_t é a saída da unidade no tempo t. Note que c_{t-1} e c_t, o estado oculto da memória no tempo $t-1$ e t, respectivamente, não são expostos.



Fonte: Adaptada de Jozefowicz, Zaremba e Sutskever (2015).

A RNN, composta de diversas camadas com unidades LSTM, foi utilizada para fundir as informações ou características extraídas de cada modalidade de entrada, sendo, portanto, responsável por realizar a fusão multimodal no modelo proposto.

Além disso, tal rede é também utilizada para identificar padrões entre tomadas adjacentes, com o objetivo de detectar o possível relacionamento semântico entre as tomadas, especificamente a mudança do assunto e, consequentemente, obtendo uma melhor segmentação em cenas. Tal resultado é fruto do pressuposto de que tomadas adjacentes, quando pertencentes a uma mesma cena, possuem alguma informação similar, ou ao menos com uma variação condizente, entre as diferentes modalidades de cada tomada. Assim sendo, graças a capacidade de aprendizagem de uma RNN e de sua célula de memória, é possível detectar que a ocorrência de determinada informação está diretamente relacionada a uma informação razoavelmente similar encontrada em alguma tomada anterior. É importante destacar que tal efeito emana diretamente da própria arquitetura desenvolvida para elementos LSTM, sendo representadas especificamente pela colaboração entre a *forget gate* f_t e a célula de memória c_t previamente descritas, e sua interação com os dados de entrada no tempo t.

O modelo específico proposto neste trabalho, consistindo de uma associação de redes convolucionais e recorrentes instanciados em duas arquiteturas diferentes, é detalhado no Capítulo 4.

2.5 Fusão multimodal

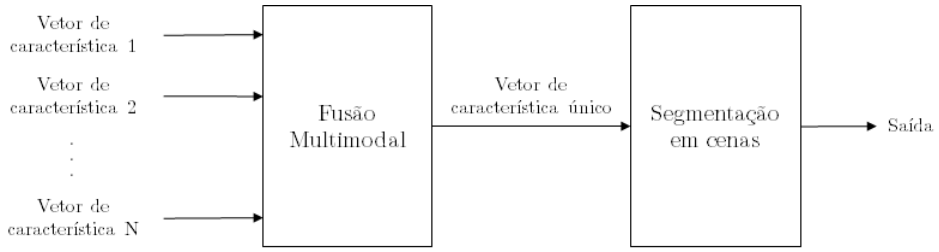
Pesquisadores da área (por exemplo, os trabalhos de Sidiropoulos *et al.* (2011), Lopes, Trojahn e Goularte (2014) e Baraldi, Grana e Cucchiara (2015a)) costumam empregar a modalidade de um vídeo como um fluxo de dados individual que pode ser processado e analisado para propósitos específicos. Em um vídeo, é comum haver três modalidades distintas: a visual, composta dos quadros do vídeo que são apresentadas a uma determinada taxa, a aural, composta

da codificação digital do áudio associado ao vídeo, além de textual, geralmente composta de legendas ou *closed captions*.

Técnicas multimodais, que usam ao menos duas modalidades diferentes, podem realizar a chamada fusão multimodal, no qual as diferentes informações são unidas em uma única representação ou resultado (ATREY *et al.*, 2010). Tal fusão pode ser adotada para propósitos diversos, sendo comum como forma de reduzir o volume de dados a serem processados ou de unir respostas diferentes obtidas por métodos ou abordagens diversas em uma única resposta. A fusão multimodal mostra-se especialmente importante para a análise de vídeos haja visto o seu carácter intrinsecamente multimídia: a completa percepção ou entendimento do assunto de um vídeo pode requerer tanto o áudio como o vídeo, por exemplo.

Em relação a tarefa de segmentação de vídeo em cenas, a fusão multimodal pode ser realizada em dois momentos distintos: antes da segmentação em cenas (fusão antecipada) ou após a mesma (fusão tardia) (ATREY *et al.*, 2010). Na fusão antecipada (do inglês *early fusion*), também conhecida como fusão de características, os vetores de características das diferentes modalidades são fundidos em um vetor de característica único logo após extraídos. Tal representação então é analisada por um algoritmo específico e a segmentação é obtida. A Figura 11 ilustra um diagrama de blocos da fusão antecipada para a segmentação em cenas dado N modalidades de entrada.

Figura 11 – Diagrama de blocos da fusão antecipada, aplicada a segmentação em cenas, dado N vetores de características de entrada, resultando em um vetor de característica único de saída, seguido do processo de segmentação em cenas.



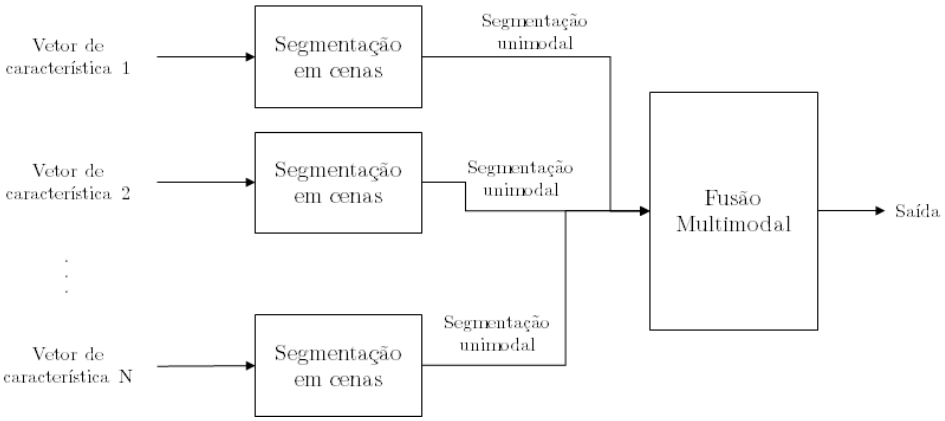
Fonte: Elaborada pelo autor.

A fusão antecipada é considerada de difícil modelagem, já que os vetores de características de entrada possuem diferentes naturezas, com diferentes dimensionalidades, representando diferentes informações extraídas da modalidade utilizada e com representações particulares e específicas para tal característica ou modalidade. Contribui para tal dificuldade, segundo Atrey *et al.* (2010), questões sobre a sincronização entre os vetores de características das modalidades, já que o mesmo evento pode ser representado em diferentes momentos em cada modalidade diferente.

Por sua vez, na fusão tardia (do inglês *late fusion*), também chamada de fusão de decisões, as características de cada modalidade são analisadas ou segmentadas individualmente. Após obter as decisões ou segmentações unimodais, seus resultados são fundidos para gerar o resultado

final da técnica. A Figura 12 ilustra um diagrama de blocos da fusão tardia para a segmentação em cenas dado N modalidades de entrada.

Figura 12 – Diagrama de blocos da fusão tardia, aplicada a segmentação em cenas, dado N vetores de características de entrada, resultando em segmentações em cenas unimodais que são providas ao método de fusão.



Fonte: Elaborada pelo autor.

Na literatura, é perceptível a ausência de uma análise sobre a fusão de modalidades relacionada a segmentação em cenas, resultando em uma falta de consenso dos pesquisadores sobre a abordagem mais adequada de fusão multimodal a ser empregada. Em tarefas relacionadas como a classificação de vídeo, por exemplo, Snoek, Worring e Smeulders (2005) desenvolveram uma técnica multimodal, vídeo-textual, avaliando o impacto do tipo de fusão multimodal no resultado. Em seu trabalho, os autores concluíram que a fusão tardia obtém, em média, melhores resultados. Nos casos na qual a fusão antecipada supera a fusão tardia, porém, a diferença é bastante significativa. Segundo a literatura, tais conclusões são, ao melhor de nosso conhecimento, ainda válidas.

Tanto a fusão tardia como a fusão antecipada possuem, devido a suas particularidades, uma série de vantagens e desvantagens quanto a outra abordagem de fusão. Algumas de suas principais diferenças são descritas a seguir:

- A fusão antecipada é considerada mais complexa e de difícil modelagem (ATREY *et al.*, 2010) já que requer a fusão e análise de características inerentemente diferentes de cada modalidade, contendo dimensionalidade e outras particularidades diversas. Já a fusão tardia, por utilizar apenas as decisões individuais, pode ser mais facilmente modelada para gerar uma segmentação única.

- Na fusão tardia, o número de execuções do algoritmo segmentador é diretamente proporcional ao número de modalidades ou características extraídas. Já a fusão antecipada requer apenas uma única execução do algoritmo segmentador, potencialmente obtendo um custo computacional menor que a fusão tardia. É importante destacar que tal particularidade é

especialmente aparente caso 1) o processo de fusão antecipada não seja demasiadamente custoso e 2) o número de modalidades/características seja significativo.

- Como consequência da particularidade menciona anteriormente, a fusão tardia apresenta um maior grau de flexibilidade para a obtenção do resultado (segmentação) final. Após as segmentações individuais (unimodais), é realizada a fusão em si, na qual as decisões serão fundidas em uma decisão ou segmentação única. Nesse processo, podem ser aplicados processos customizados de otimização com o intuito de melhor evidenciar resultados desejáveis e, ao mesmo tempo, ignorar ou suprimir resultados incorretos. Um procedimento similar na fusão antecipada, por sua vez, pode requer uma reformulação do processo de construção do vetor de característica único e do próprio algoritmo segmentador, sendo, portanto, potencialmente impraticável na maioria dos casos.

- A fusão tardia permite um melhor distribuição de processamento em sistemas distribuídos. Como a fusão ocorre apenas quando todas as decisões individuais foram obtidas, a análise de cada modalidade pode ser realizada paralelamente, potencialmente em dispositivos ou computadores diferentes. A fusão antecipada, porém, requer que todas as modalidades estejam disponíveis no momento da criação do vetor de característica multimodal.

- A adição de uma nova modalidade ou característica a ser processada é, potencialmente, mais simples na abordagem de fusão tardia. Como o processo de fusão ocorre sobre os resultados ou segmentações unimodais, a adição de uma nova modalidade ou característica não incorre em grandes alterações ao algoritmo de fusão. Por sua vez, na fusão antecipada, tanto o procedimento de fusão em um vetor de característica único, como também o de segmentação em si, podem requerer alterações mais complexas para suportar tal configuração.

É importante destacar que, embora a fusão tardia seja de mais fácil modelagem, dificulta ou impossibilita a análise do correlacionamento de informações entre as diferentes modalidades ou características. Por exemplo, caso a modalidade visual indique um evento X e a modalidade aural indique Y, os algoritmos de segmentação individuais devem determinar a existência da transição de cenas tendo conhecimento apenas dos eventos X (visual) ou Y (aural). A fusão antecipada, por sua vez, têm acesso direto às características dos eventos X e Y, podendo então obter um resultado completamente diverso do que as decisões individuais da fusão tardia poderiam obter. É importante mencionar, porém, que tal processo pode ser tanto benéfico, no qual o relacionamento entre as modalidades resulta em uma segmentação mais próxima da desejada, como prejudicial, no qual o relacionamento entre as modalidades adiciona ruído a técnica e dificulta a detecção da segmentação em cenas. Embora ainda considerado um problema em aberto, a seleção criteriosa de características (Jesus; Araújo; Canuto, 2016; Hildebrandt, 2015) ou a filtragem de características como o *Random Sample Consensus* (RANSAC) (FISCHLER; BOLLES, 1981) podem evitar tal limitação.

2.6 Avaliação de eficácia

Um importante aspecto do desenvolvimento e aperfeiçoamento de técnicas de segmentação em cena é o como tais técnicas são avaliadas para determinar os limites individuais de cada abordagem. Nesse sentido, a avaliação de sua eficácia consiste em aplicar a técnica proposta para segmentar um vídeo e comparar o resultado obtido com uma base confiável (do inglês *ground truth*), composta de vídeos e suas respectivas anotações, medindo o quão próximo a segmentação obtida está da segmentação considerada ideal. Assim, a Subseção 2.6.1 apresenta algumas considerações quanto a bases de vídeo confiáveis para a segmentação em cenas. Por sua vez, as métricas amplamente utilizadas por pesquisadores da área são apresentadas e discutidas na Subseção 2.6.2.

2.6.1 Base confiável

Uma base confiável, como mencionado previamente, é fundamental para a avaliação e comparação de uma técnica de segmentação em cenas frente a outras propostas. A ausência de uma base confiável comum, usada por diferentes técnicas, dificulta a comparação entre técnicas já que, nesse caso, ter-se-ia diferentes particularidades em cada base, como definição de cena utilizada, qualidade dos dados de entrada, domínio adotado, entre outros.

Conforme mencionado em diversos trabalhos reportados na literatura (FABRO; BÖSZÖRMENYI, 2013), há poucas bases de dados publicas adequadas para a tarefa de segmentação de vídeo em cenas. Uma base confiável adequada é formada por, pelos menos, 1) os dados de entrada necessários, como um conjunto de características ou o próprio vídeo e 2) algum tipo de descrição das cenas e/ou de suas transições. Em geral, bases já propostas para a tarefa apresentam limitações importantes tais como:

- Não são publicamente acessíveis: diversas bases reportadas na literatura não são acessíveis, no sentido que não há meio de pesquisadores ou interessados em geral obterem cópia da base para seus propósitos. Algumas bases de vídeos antigas da TRECVid[1], por exemplo, requerem a requisição e o recebimento de mídias físicas contendo os dados.

- Não possuem o *ground-truth*: algumas bases, como a CCV *Dataset*[2] disponibilizam o vídeo para *download*, mas, ou não são voltadas para a segmentação em cenas, ou não disponibilizam a sua respectiva base confiável contendo anotações das respectivas cenas.

- Não são adequadas para a tarefa: para a segmentação em cenas, é recomendável haver uma quantidade relevante de vídeos com qualidade de vídeo aceitável, de maneira a exercitar adequadamente a eficácia de cada técnica proposta. Bases públicas como a TRECVid, por exemplo, apresentam ruídos de codificação e de conversão analógico-digital que podem

[1] <https://trecvid.nist.gov/past.data.table.html>
[2] <http://www.ee.columbia.edu/ln/dvmm/CCV/>

impactar negativamente a eficácia de técnicas de segmentação em cenas. Adicionalmente, é necessário a presença de vídeos com número significativo de tomadas e cenas, algo que torna a inadequada a CCV *Dataset*, formada majoritariamente de vídeos com apenas uma única tomada, por exemplo.

- Possuem restrições de licenciamento: diversas das bases adotadas são formadas por vídeos com restrições de licenciamento e não podem ser disponibilizadas publicamente. Exemplos de tal restrição são as bases de vídeos de filmes utilizadas nos trabalhos de (RASHEED; SHAH, 2003), (CHASANIS; KALOGERATOS; LIKAS, 2009) e (SIDIROPOULOS *et al.*, 2011), entre outros.

O resultado de tais limitações, como relatado por Fabro e Böszörmenyi (2013), é que a maioria dos trabalhos seminais de segmentação em cenas utilizam bases de dados customizadas, não disponíveis, dificultando a comparação entre técnicas diferentes, o que pode refletir em um viés nos resultados obtidos e comprometer a reprodutibilidade da técnica reportada. Recentemente, porém, duas bases de dados foram disponibilizadas publicamente, como a BBC *Dataset* (BARALDI; GRANA; CUCCHIARA, 2015a) e o IBM *Open Video Scene Detection* (OVSD) (ROTMAN; PORAT; ASHOUR, 2016), contendo tanto o vídeo em si como sua correspondente anotação de cenas. Tais bases, portanto, podem ser facilmente utilizadas para a avaliação de técnicas de segmentação em cenas por meio de métricas preestabelecidas.

A BBC *Dataset* é uma base confiável disponibilizada[3] por Baraldi, Grana e Cucchiara (2015a), contendo as anotações em tomadas e cenas de um conjunto de onze vídeos documentários da série BBC *Planet Earth*[4]. As informações acerca de tal base confiável são apresentada na Tabela 1.

A BBC *Dataset* é composta de vídeos altamente padronizados, com pequena variação entre o número de tomadas, cenas e duração total. É perceptível a baixa duração média tanto das tomadas (6 segundos) como das cenas (48 segundos). Além disso, a base confiável contém um número elevado de cenas pequenas, com até 3 tomadas de duração, e, proporcionalmente, um menor número de cenas grandes, com pelo menos 10 tomadas de duração. Nesse sentido, a Tabela 2 descreve o tamanho das cenas da BBC *Dataset*.

Tal base é considerada adequada para a avaliação de técnicas de segmentação em cenas por possuir conteúdos bastante variados, como diferentes ambientes representados em cada vídeo, assim como diferentes objetos (flora e fauna) sendo representados e mencionados durante a duração de cada vídeo. Outras particularidades que a torna uma base adequada a tarefa é o fato de possuir vídeos completos e de considerável duração, além de um número significativo de

[3] A BBC *Dataset* pode ser requisitada no endereço <http://imagelab.ing.unimore.it/imagelab/page.asp?IdPage=5>

[4] <https://www.bbc.co.uk/programmes/b006mywy>

Tabela 1 – Nome do episódio, duração e número de tomadas e cenas da base confiável BBC Planet Earth disponibilizado por Baraldi, Grana e Cucchiara (2015a).

Nome	Duração (hh:mm:ss)	N. tomadas	N. cenas
From Pole to Pole	49:15	450	66
Mountains	48:04	395	53
Ice Worlds	49:17	425	62
Great Plains	49:03	473	71
Jungles	49:14	461	65
Seasonal Forests	49:19	529	65
Fresh Water	41:17	533	62
Ocean Deep	49:14	418	53
Shallow Seas	49:14	367	61
Caves	48:55	393	57
Deserts	48:59	469	55
Total	08:51:51	4913	670

Tabela 2 – Número de tomadas de duração das cenas e sua proporção na base confiável da BBC *Dataset*.

Nome	1 tomada	até 3 tomadas	ao menos 10 tomadas
From Pole to Pole	7 (11%)	17 (26%)	8 (12%)
Mountains	6 (11%)	16 (30%)	13 (25%)
Ice Worlds	4 (6%)	17 (27%)	14 (23%)
Great Plains	4 (6%)	19 (27%)	12 (17%)
Jungles	9 (14%)	26 (40%)	12 (18%)
Seasonal Forests	3 (5%)	15 (23%)	14 (22%)
Fresh Water	4 (6%)	13 (21%)	16 (26%)
Ocean Deep	2 (4%)	18 (34%)	10 (19%)
Shallow Seas	7 (11%)	23 (38%)	12 (20%)
Caves	4 (7%)	17 (30%)	11 (19%)
Deserts	8 (15%)	15 (27%)	13 (24%)
Média	5 (9%)	18 (29%)	12 (20%)

tomadas e cenas. Características estas que justificam a sua utilização na avaliação realizada neste trabalho, apresentada no Capítulo 5.

Por sua vez a base confiável IBM Open Video Scene Segmentation, proposta por Rotman, Porat e Ashour (2016) contém cinco curtas animados e um filme de longa duração, disponíveis publicamente. Assim como na BBC *Dataset*, tanto a segmentação em cenas como a de tomadas estão disponíveis, além dos vídeos em si no site de seus respectivos criadores. Tal base confiável, porém, não foi empregada na avaliação realizada pois: contém vídeos na sua maioria com pequena duração, que resulta em uma base com baixo número de tomadas e cenas; contém vídeos sem qualquer tipo de narração ou fala, contando apenas com trilha sonora de fundo, o que influencia a avaliação da eficácia dado que a proposta utiliza tanto informações aurais como textuais. Finalmente, tal base não foi utilizada para comparar técnicas de segmentação consideradas em estado da arte.

2.6.2 Métricas

É necessário o uso de uma métrica preestabelecida para medir o quão distante o resultado de uma técnica de segmentação está da segmentação ideal. Nesse sentido, a literatura reporta diversas métricas diferentes, sendo as mais utilizadas pelos pesquisadores descritas nesta seção.

A Precisão (do inglês *precision* - *P*) e a Abrangência (do inglês *Recall* - *R*) (RIJSBERGEN, 1979), originárias da área de recuperação de informação, são duas métricas bem conhecidas. No contexto de segmentação em cenas, como cada tomada é ou não é de transição, o conjunto resposta do segmentador inclui os seguintes casos: Verdadeiro Positivo (vp), quando uma tomada apresentada como sendo de transição corresponde de fato a uma transição na base confiável; Falso Positivo (fp), quando uma tomada apresentada como sendo de transição não corresponde a uma transição de fato; Verdadeiro Negativo (vn), é uma tomada não apresentada na resposta do segmentador e que de fato não é de transição; Falso Negativo (fn), é uma tomada não apresentada pelo segmentador mas que, na realidade, é de transição. Assim, Precisão P e Abrangência R são definidas como:

$$P = \frac{vp}{vp + fp} \tag{2.16}$$

$$R = \frac{vp}{vp + fn} \tag{2.17}$$

A Precisão e Abrangência avaliam duas características diferentes de um segmentador qualquer. Caso um algoritmo obtenha uma elevada Precisão, significa dizer a maioria das transições detectadas pelo algoritmo são transições verdadeiras, ou seja, o número de falsos positivos obtido é baixo. Mas não implica que todas as transições verdadeiras foram detectadas. Por outro lado, caso o algoritmo obtenha uma elevada Abrangência, significa que a maioria das transições que existem na base confiável foram detectadas, indicando um baixo número de falsos negativos. Mas não implica que somente transições verdadeiras foram detectadas.

Como tanto P e R medem diferentes aspectos da segmentação em si, um modo de agrupá-los em um valor único é por meio da F-measure ou F-score (RIJSBERGEN, 1979). A F-measure é uma média harmônica entre os valores P e R que apresenta como vantagem o fato de ser mais conservadora que a média aritmética simples entre P e R, podendo inclusive dar maior prioridade tanto a P quanto a R em variantes tais como $F_{0.5}$ ou F_2. A métrica F-measure F_β é definida como:

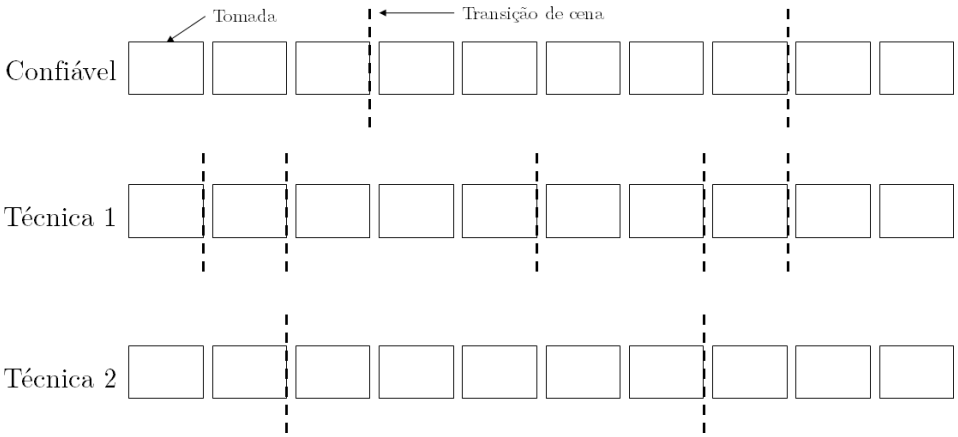
$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot A}{(\beta^2 \cdot P) + A} \tag{2.18}$$

Onde β é usado para dar prioridade para a Precisão (valores abaixo de um) ou para a Abrangência (valores acima de um). Caso $\beta = 1$ (F_1), não há prioridade qualquer, sendo a mesma doravante chamada de F_{PR}, definida como:

$$F_{PR} = 2 \cdot \frac{P \cdot A}{P + A} \tag{2.19}$$

Embora amplamente utilizadas, principalmente em trabalhos seminais na área (FABRO; BÖSZÖRMENYI, 2013), as métricas de Precisão e Abrangência possuem limitações quanto a segmentação em cenas. Nesse sentido, a Figura 13 ilustra duas técnicas que obtém, para uma determinada segmentação confiável com duas transições, diferentes segmentações. Note que os valores de P, R e F_{PR} da **Técnica 1** (20%, 50% e 28% respectivamente), são superiores a **Técnica 2** (0%, 0% e 0% respectivamente), embora intuitivamente a segunda tenha obtido uma segmentação mais próxima da desejada (apenas deslocada por uma tomada).

Figura 13 – Ilustração de uma segmentação em cenas confiável (desejada) e de duas técnicas hipotéticas para um determinado video confiável com 10 tomadas de duração. Os retângulos representam as tomadas e a as linhas verticais denotam as transições de cenas.



Fonte: Elaborada pelo autor.

Por tal motivo, diversos pesquisadores costumam adotar uma janela de tolerância, normalmente baseada em um número de quadros de vídeo, tempo ou número de tomadas (BARALDI; GRANA; CUCCHIARA, 2015b). Por exemplo, Rasheed e Shah (2003) especificam que uma transição de cena é considerada correta se estiver até dez segundos da transição confiável, enquanto que Hanjalic, Lagendijk e Biemond (1999) usaram três tomadas de tolerância. Tal prática, porém, causa inconvenientes tais como a dificuldade de comparar técnicas que usaram janelas de tolerância diferentes, além de ocultar pequenas diferenças que diferentes técnicas poderiam gerar. Fabro e Böszörmenyi (2013) citam como outra possível limitação das métricas a dificuldade de medir quando uma cena é parcialmente detectada, quando o início e fim da cena não estão alinhados.

Nesse sentido, Vendrig e Worring (2002) definiram duas métricas específicas para a segmentação em cenas, conhecidas como Cobertura (do inglês *Coverage* - C) e Transbordamento (do inglês *Overflow* - O). A Cobertura procura medir quantas tomadas pertencentes à mesma cena foram corretamente agrupadas. Já o Transbordamento mede quantas tomadas, mesmo não pertencentes a mesma cena, foram incorretamente agrupados em uma mesma cena. As medidas

de Cobertura (C) e Transbordamento (O) de uma determinada cena confiável são definidas como:

$$C(x_t) = \frac{\max_{j=0\ldots n} \#(y_j)}{\#(x_t)} \tag{2.20}$$

$$O(x_t) = \frac{\sum_{j=0}^{n} \#(y_j \backslash x_t) \cdot \min(1, \#(y_j \cap x_t))}{\#(x_{t-1}) + \#(x_{t+1})} \tag{2.21}$$

Sendo x_t o conjunto de tomadas que forma a cena confiável de índice t, y_j o conjunto de tomadas que forma a cena de índice j obtida pelo algoritmo segmentador, $\#(x)$ o operador que retorna a quantidade de tomadas de uma determinada cena x e $\#(y_j \backslash x_t)$ o número de tomadas da cena detectada que não se encontram também na cena confiável (ou, em outras palavras, a operação de diferença entre y_j e x_t). É importante ressalvar que, ao contrário das métricas P, R e C, no qual o valor desejado é igual a 1 (100%), o valor desejado de Transbordamento (O) é igual a zero (0%), que corresponde ao caso de que nenhuma tomada de uma cena adjacente foi erroneamente agrupada na cena detectada.

Para obter o valor de C ou O de um dado vídeo V, a média ponderada é calculada de acordo com o número de tomadas do vídeo. Assim, sendo $\#(V_\sigma)$ o número total de tomadas e $\#(V_{x_j})$ o número de cenas confiáveis do vídeo V, o cálculo da Cobertura $C(V)$ e do Transbordamento $O(V)$ é dado por:

$$C(V) = \sum_{t=0}^{\#(V_{x_j})-1} C(x_t) \cdot \frac{\#(x_t)}{\#(V_\sigma)} \tag{2.22}$$

$$O(V) = \sum_{t=0}^{\#(V_{x_j})-1} O(x_t) \cdot \frac{\#(x_t)}{\#(V_\sigma)} \tag{2.23}$$

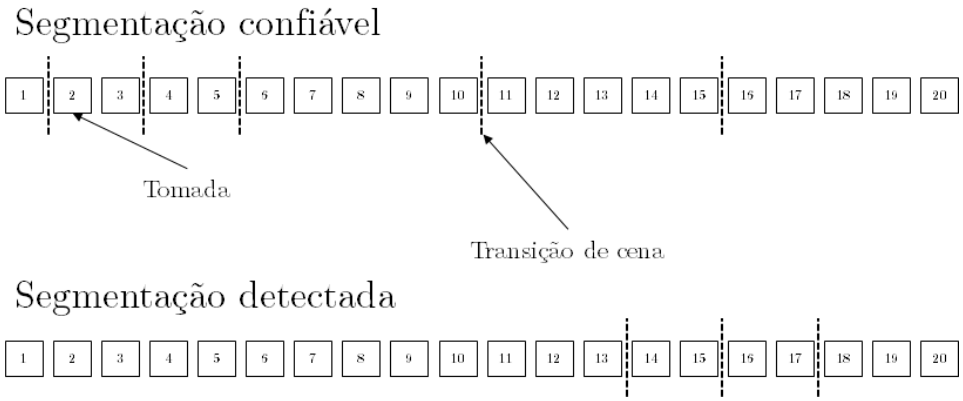
Assim como no caso de Precisão e Abrangência, os valores de Cobertura e Transbordamento podem ser agrupados por meio da média harmônica F-measure, doravante representado como F_{CO}. É importante ressalvar que o valor do cálculo de Transbordamento utilizado na métrica F_{CO} é igual a $1 - O$, visto a característica do valor desejado de O ser igual a zero.

Uma das principais vantagens das medidas de Cobertura/Transbordamento sobre a Precisão/Abrangência é não requerer a especificação de um parâmetro de tolerância, que pode interferir na interpretação dos resultados da técnica avaliada, facilitando inclusive a comparação com outras técnicas desenvolvidas. Outro benefício é a de que a métrica avalia o *quanto* uma cena foi detectada, sendo, portanto, mais robusta a cenas parcialmente detectadas (FABRO; BÖSZÖRMENYI, 2013).

Além disso, a formulação da métrica de Transbordamento definida anteriormente contém uma limitação importante caso a segmentação a ser avaliada possua alto índice de subsegmentação (do inglês *undersegmentation*), quando uma cena detectada engloba diversas cenas

confiáveis (que deveriam ser detectadas), resultando em valores inválidos. Para ilustrar tal limitação, considere um vídeo formado de 20 tomadas com 6 cenas confiáveis (5 transições) e o resultado obtido por uma técnica de segmentação com 4 cenas (3 transições) sobre o mesmo vídeo. A Figura 14 ilustra o vídeo e as duas segmentações, confiável e detectada, mencionados.

Figura 14 – Exemplos hipotéticos de uma segmentação confiável e detectada de um trecho de vídeo de 20 tomadas. Os retângulos representam as tomadas e a as linhas verticais denotam as transições de cenas.



Fonte: Elaborada pelo autor.

Embora a segmentação detectada obtenha uma Cobertura válida de 80%, já que apenas 4 tomadas (20%) da segmentação confiável não são cobertas pela respectiva cena detectada (tomadas 14 a 17), o valor de Transbordamento é de 135%, um valor claramente inválido que excede o intervalo válido da métrica. Consequentemente, nesse caso, o valor obtido da métrica F_{CO} (-124%) também é inválido.

Para evitar tal problema, Han e Wu (2011) apresentam uma formulação alternativa para a métrica de Transbordamento, no qual é normalizado o valor para o intervalo válido $[0,1]$. Tal métrica, doravante chamada de NO (*New Overflow*), é definida como:

$$NO(x_t) = 1 - \frac{\#(x_t)}{\sum_{j, x_t \cap y_j \neq \emptyset} \#(y_j)} \tag{2.24}$$

A principal melhoria da formulação da métrica NO sobre a métrica O é o denominador da divisão, formada pela soma do número de tomadas de todas as cenas detectadas que possuem alguma intersecção com a cena confiável sendo analisada. Tal modificação significa que o denominador sempre será maior ou igual (melhor caso) que o numerador o que, associado com a subtração da equação, garante que o valor NO será um valor válido no intervalo $[0,1]$. Assim como na métrica O, é calculada a média harmônica F-measure de maneira semelhante a métrica F_{CO}, usando o valor $1 - NO$, sendo a mesma doravante representada como F_{CNO}. Graças a tal melhoria, no exemplo hipotético ilustrado na Figura 14, o valor obtido de NO é de cerca de 53% (0.53) e seu valor F_{CNO} é igual a cerca de 59% (0.59), ambos valores válidos do intervalo $[0,1]$.

Além das previamente mencionadas, outras métricas são reportadas na literatura para a segmentação em cenas. Por exemplo, Sidiropoulos *et al.* (2012) definem uma métrica chamada de *Differential Edit Distance* (DED), baseada na distância de edição diferencial, no qual é calculado o número de operações para que uma dada segmentação obtida se assemelhe a segmentação desejada. Vinciarelli e Favre (2007), por sua vez, definem uma métrica chamada de *purity*, baseada no cálculo do tempo de intersecção entre a segmentação confiável e a detectada. Tais métricas, porém, até onde se sabe, não foram utilizadas por outros autores em trabalhos posteriores.

2.7 Discussões sobre o capítulo

Neste capítulo foram apresentados os conceitos fundamentais necessários para a compreensão da proposta deste trabalho, incluindo a definição de vídeo digital, o processo geralmente adotado de segmentação em cenas e a extração de características. Além disso, a Aprendizagem Profunda, as abordagens de fusão multimodal e questões sobre a avaliação da eficácia de técnicas de segmentação em cenas também foram discutidas.

Um vídeo digital pode ser temporalmente segmentado em quadros, tomadas e cenas. Destas, a segmentação em cenas é especialmente desafiadora devido ao fato de a mudança de assunto, que sinaliza a transição de cena, ser altamente subjetiva, sendo, portanto, de alto nível semântico. Neste trabalho foi adotada a definição usual de cena, formada por tomadas adjacentes semanticamente relacionadas, sendo esta uma definição flexível, capaz de descrever cenas nas mais diversas situações. Finalmente, por ser uma mídia não-estruturada e de alto volume, é costumeiramente aplicado o processo de extração de características sobre um vídeo digital, efetivamente reduzindo o volume de dados a serem posteriormente processados. Tal processo, assim como a descrição de três extratores de características amplamente adotados na literatura, também foi discutido neste capítulo.

A Aprendizagem Profunda, uma área de pesquisa relacionada à aprendizagem de máquina e inteligência artificial, dispõe de uma ampla gama de abordagens recentes que podem beneficiar diversas tarefas multimídia, como a segmentação de vídeo em cenas. As redes convolucionais, popular abordagem da Aprendizagem Profunda, por exemplo, trouxeram melhorais significativas na extração de características de imagens, efetivamente superando humanos em tarefas específicas como a classificação de imagens (HE *et al.*, 2015). As redes recorrentes, por sua vez, possuem um grande potencial devido a sua capacidade de aprender informações sequenciais. Uma rede neural convolucional, com grande potencial para extração de características adequadas de diferentes modalidades, associada a uma rede recorrente capaz de aprender padrões temporais e de realizar uma fusão multimodal significativa, utilizando informações complementares para determinar temporalmente mudanças significativas no assunto, pode resultar em avanços perceptíveis em tarefas multimídia como a segmentação em cenas.

As duas principais abordagens de fusão multimodal, a fusão antecipada e a fusão tardia, foram discutidas na Seção 2.5. Em geral, as técnicas multimodais desenvolvidas, devido a dificuldades de modelagem do problema, não fazem o uso das diferentes particularidades de cada abordagem de fusão, optando por uma abordagem de fusão atrelado as características usadas ou a técnica desenvolvida. Em técnicas tradicionais, que não utilizam Aprendizagem de Máquina, a prevalência é pela fusão tardia das modalidades, principalmente devido a dificuldades de fundir características de diferentes naturezas. Já no caso da Aprendizagem Profunda, que oferece melhor suporte a características diferentes, é preferida a fusão antecipada. O desenvolvimento de um método que permita o uso flexível de diferentes abordagens de fusão multimodal, tanto antecipada como tardia, pode resultar em melhorias significativas em tarefas como a segmentação de vídeo em cenas.

Por fim, discussões sobre aspectos da avaliação da eficácia de técnicas de segmentação em cenas foram apresentadas. Um importante aspecto da avaliação de técnicas é a falta de bases de vídeos públicas reportadas por diversos pesquisadores, principalmente em artigos seminais da área. Tal fato tem por consequência a dificuldade de comparar técnicas diferentes. Entretanto, algumas bases recentes adequadas para a tarefa de segmentação em cenas são reportadas na literatura, tal como a BBC *Dataset*, formada por onze vídeos documentários sobre diferentes habitats, utilizado neste trabalho para avaliar as duas arquiteturas desenvolvidas. Outro aspecto da avaliação é a escolha de métricas adequadas para medir a distância da segmentação obtida pela técnica com a segmentação desejada. Nesse sentido, foram apresentadas as principais métricas adotadas por pesquisadores da área, incluindo métricas tradicionais como a Precisão e Abrangência, oriundas da área de recuperação de informação, assim como métricas específicas da segmentação em cenas como a Cobertura e o Transbordamento. A Precisão e Abrangência, conforme reportado em trabalhos recentes, possuem diversas limitação que as tornam inadequadas para a segmentação em cenas, como a incapacidade de distinguir a magnitude do erro de uma transição incorretamente detectada. As métricas de Cobertura e Transbordamento, por sua vez, se mostram muito mais robustas e adequadas para a tarefa, sendo geralmente adotada por trabalhos recentes da área. O desenvolvimento de novas métricas, porém, indicam que não há um consenso, entre os pesquisadores, sobre o como avaliar a segmentação em cenas. É comum, por exemplo, que diferentes métricas deem diferentes veredictos sobre uma mesma segmentação. Para facilitar comparações com técnicas futuras, as principais métricas amplamente adotadas na área foram utilizadas na avaliação apresentada no Capítulo 5.

CAPÍTULO

3

TRABALHOS RELACIONADOS

A literatura reporta diversos trabalhos contendo discussões de técnicas desenvolvidas para a segmentação em cenas nas últimas décadas. Manzato, Fortes e Goularte (2006), por exemplo, realizaram uma discussão acerca de diversos aspectos da segmentação de vídeo, com foco na apresentação de diversas abordagens seminais, empregadas em tarefas como a segmentação em tomadas, cenas e de objetos. Vendrig e Worring (2002), por sua vez, reportam a inadequação das métricas de Precisão e Abrangência para a avaliação da segmentação em cenas. Além de propor duas novas métricas, a Cobertura e o Transbordamento, realizam uma análise da eficácia de algumas das abordagens mais utilizadas à época em diversas bases de dados. Petersohn (2008) detalha diversas particularidades pertinentes a segmentação em cenas, como a escolha das características, método de detecção das transições de cenas e escolha de parâmetros de algumas técnicas populares. Mais recentemente, Fabro e Böszörmenyi (2013) ampliaram a discussão de Vendrig e Worring (2002), classificando as técnicas de acordo com as modalidades que empregam, discutindo ainda diversos aspectos da tarefa tal como a dificuldade de comparação entre técnicas, métricas de comparação geralmente adotadas, entre outros.

Usando como base tais discussões, além de outras técnicas mais recentes reportadas na literatura, foram selecionados um conjunto de abordagens seminais e outro que representa o estado da arte na área de acordo com os resultados obtidos em avaliações usando bases de dados adequadas. As técnicas seminais selecionadas, discutidas na Seção 3.1, apresentam abordagens maduras, amplamente utilizadas como referência no desenvolvimento de técnicas mais recentes. Por sua vez, as técnicas em estado da arte selecionadas apresentam os melhores resultados reportados em suas respectivas abordagens, incluindo técnicas unimodais, multimodais e multimodais baseadas em Aprendizagem Profunda. As técnicas relacionadas em estado da arte são discutidas na Seção 3.2 .

3.1 Técnicas de segmentação seminais

Dentre as diversas técnicas reportadas na literatura, desenvolvidas sem um domínio específico, estão presentes técnicas inovadoras ou que obtiveram, em seu tempo, avanços significativos na tarefa proposta, embora tenham sido superadas em eficácia por abordagens mais recentes. A análise de tais abordagens indica as direções mais promissoras que influenciaram técnicas atuais, algumas consideradas em estado da arte e apresentadas na Seção 3.2. Nesse sentido, as técnicas seminais geralmente mencionadas e relacionadas ao trabalho aqui proposto são discutidas nas subseções a seguir.

3.1.1 Técnica baseada em grafos

Yeung, Yeo e Liu (1998) propuseram uma das primeiras técnicas unimodais de segmentação de vídeo em cenas baseada em grafos, obtidas pelo cálculo de similaridade visual, com restrições temporais. Na técnica proposta, após a segmentação do vídeo de entrada em tomadas e seleção de um conjunto variável de quadros-chaves (do inglês *key frames*) para cada uma, as tomadas são agrupadas de acordo com a similaridade visual com restrições temporais. Tais restrições são usadas para impedir que tomadas visualmente similares, mas distantes temporalmente no vídeo, sejam agrupadas. A partir daí, é construído o chamado grafo de transição de cena (do inglês *Scene Transition Graph* - STG), criado originalmente para a representação de cenas e sua navegação.

O STG é construído ao considerar cada grupo como um vértice, no qual haverá uma arresta direcionada entre diferentes grupos se houver duas tomadas que sejam temporalmente adjacentes. As transições em cenas em si são detectadas ao identificar as arrestas de corte, quando a sua remoção implica como resultado um subgrafo não conectado. O algoritmo de criação do grafo de transição de cenas (STG) é descrito no Algoritmo 1, considerando como entrada um conjunto de tomadas $S = s_1, ..., s_n$, um histograma $H = h_1, ..., h_n$, o tamanho de janela ω e um limiar de agrupamento δ. A saída do algoritmo STG é uma lista $B = b_1, ..., b_n$ de transições de cenas.

Como pode ser percebido no Algoritmo 1, o algoritmo para identificar as transições de cenas por meio da STG dependem de dois parâmetros distintos: ω, utilizado como restrição temporal para o cálculo de distância entre duas tomadas quaisquer e δ, o limiar máximo de distância entre grupos usado como critério de parada do agrupamento. Como percebido pelos autores, a eficácia da técnica depende da seleção correta de tais parâmetros, já que, por exemplo, um valor ω demasiadamente elevado pode resultar no agrupamento incorreto de muitas tomadas, enquanto que um valor demasiadamente pequeno pode afetar o agrupamento de tomadas da mesma cena.

A abordagem baseada em grafos desenvolvida foi amplamente utilizada como inspiração para diversas técnicas posteriores. Rasheed e Shah (2005), por exemplo, criam um grafo não-

Algoritmo 1 – Algoritmo para criação do grafo de transição de cena (STG)

função STG(S, H, ω, δ)
    Faça a conexão entre as S tomadas usando a distância do cosseno entre seus correspondentes H histogramas. Se um par de tomadas estiver ω ou mais tomadas distantes, considere a distância como infinita. Pare quando a distância entre-grupos entre todos os grupos for maior que δ e adicione os grupos restantes ao conjunto C.
    Crie o grafo $G = V, E$
    $V \leftarrow$ índices de C
    Insere as arestas $e = (a, b)$ em E, sendo $a, b \in V$ para todos as tomadas $s_i \in c_a$ e $s_{i+1} \in c_b$, com $s_i, s_{s+1} \in S$ e $c_a, c_b \in C$
    Encontra e remova todas as arestas de corte em G usando a busca em profundidade. Adicione os subgrafos resultantes para o conjunto R.
    Crie a lista vazia B
    enquanto houver grafo $r \in R$ **faça**
        FS(r) $\leftarrow$ Índice da primeira tomada representada por r
        LS(r) $\leftarrow$ Índice da última tomada representada por r
        Insere FS(r) e LS(r) em B
    fim enquanto
    Ordena B
    retorna B
fim função

direcionado valorado chamado de Grafo de Similaridade de Tomadas (do inglês *Shot Similarity Graph*) baseado na similaridade entre histogramas e com restrições temporais que, por meio do algoritmo de separação de grafos de cortes normalizados proposto por Shi e Malik (2000), é usado para detectar as transições de cenas. Sidiropoulos *et al.* (2011), por sua vez, desenvolveram uma técnica em estado da arte por meio da fusão multimodal entre STGs criados para cada característica utilizada. Tal técnica é detalhada na Subseção 3.2.2.

3.1.2 Técnica baseada na coerência visual

Seguindo uma abordagem diferente da proposta por Yeung, Yeo e Liu (1998), Rasheed e Shah (2003) propuseram analisar a coerência visual entre tomadas adjacentes usando, para isso, histogramas de cor e uma medida de "quantidade de movimento". Na abordagem proposta, a similaridade visual entre dois quadros f^x e f^y ($D(f^x, f^y)$) é calculada como a intersecção de histogramas, definida como a soma do menor valor de cada *bin* do histograma, conforme a Equação 3.1.

$$D(f^x, f^y) = \sum_{b \in bins} \min(H_x[b], H_y[b]) \tag{3.1}$$

Onde H_x e H_y são os histogramas dos quadros f^x e f^y respectivamente e b o índice do vetor do histograma dividido em *bins*.

O primeiro passo da técnica proposta consiste na segmentação em tomadas, baseada na comparação da intersecção de histogramas entre quadros adjacentes a um limiar predefinido T_{color}. Ou seja, uma transição de tomadas é identificada quando a seguinte equação é verdadeira:

$$D(f^x, f^{x-1}) < T_{color} \tag{3.2}$$

Após encontradas as transições de tomadas, é selecionado um conjunto K de quadros-chave para cada tomada, formado pelo quadro mediano (que se localiza no centro da tomada) e todos os demais que possuírem intersecção de histogramas abaixo de um limiar predefinido. Tal estratégia tem por objetivo a seleção de um número variável de quadros conforme as particularidades da tomada em análise.

Com um conjunto de quadros-chave para cada tomada, é calculado o grau de similaridade visual entre tomadas adjacentes, chamado de coerência entre tomadas (do inglês *Shot Coherence*), definido como o maior valor da intersecção de histogramas entre histogramas dos quadros-chave dois-a-dois. A coerência entre tomadas i e j, denotada como SC_i^j, é dada por:

$$SC_i^j = \max_{f^x \in K_i, f^y \in K_j} (D(f^x, f^y)) \tag{3.3}$$

Onde K_i e K_j são os conjuntos de quadros chave das tomadas i e j, e $D(f^x, f^y)$ é a função que retorna a intersecção de histogramas entre quadros f^x e f^y. Para calcular a similaridade de uma tomada com as N tomadas anteriores, evitando casos de mínimo/máximo locais na comparação dois-a-dois, é estimada a chamada *Backward Shot Coherence* (BSC), definida como:

$$BSC_i = \max_{i \leq k \leq N} (SC_i^{i-k}) \tag{3.4}$$

Por fim, após o cálculo do valor BSC para cada tomada, as transições de cenas podem ser detectadas encontrando os mínimos locais do valor BSC. Tais pontos, que apresentam baixa similaridade com as N tomadas anteriores, são consideradas dissimilares o bastante para indicarem o começo de uma nova cena.

Conforme percebido pelos autores, o algoritmo apresenta uma alta taxa de sobre-segmentação (do inglês *oversegmentation*) em trechos de alta movimentação, que geralmente englobam tomadas com baixa similaridade visual e de baixa duração. Para reduzir tal sobre-segmentação, os autores propõem, o cálculo da "quantidade de movimento". Na proposta, os vetores de movimento presentes do fluxo de dados MPEG são utilizados para calcular a magnitude do movimento de uma tomada que, dividida pelo número de quadros da tomada, resulta no valor de quantidade de movimento. Caso duas potenciais transições de cenas (tomadas) possuírem valores de quantidade de movimento acima de um limiar pré-fixado, tais transições são removidas, efetivamente reduzindo a sobre-segmentação.

Infelizmente, tanto a base confiável utilizada para a avaliação de eficácia quanto a implementação da técnica não estão disponíveis publicamente, impossibilitando o seu uso na avaliação realizada neste trabalho.

Abordagens similares, que procuram calcular e identificar o correlacionamento de uma tomada com as tomadas anteriores, foram desenvolvidas posteriormente. Chasanis, Likas e Galatsanos (2009), por exemplo, desenvolveram uma técnica unimodal em estado da arte na qual as tomadas são rotuladas e agrupadas. Por meio da identificação de sequências de rótulos, usando um algoritmo de alinhamento de sequências, as transições de cenas podem ser identificadas. Tal técnica é detalhada na Subseção 3.2.1. Por sua vez, Trojahn e Goularte (2013) propuseram diversas melhorias para a abordagem original, como a introdução de um termo de memória com o intuito de progressivamente reduzir a influência de tomadas mais distantes a tomada sendo analisada no cálculo da correlação, além da adoção da análise da "quantidade de movimento" baseada no fluxo óptico (do inglês *optical flow*), uma abordagem consideravelmente mais flexível e que suporta vídeos de qualquer formato, uma limitação importante presente na abordagem de Rasheed e Shah (2003).

3.1.3 Técnica baseada em palavras visuais

Uma outra abordagem para a segmentação em cenas é baseada no modelo *Bag of Visual Words* (BoVW), inspirado no *Bag of Words* (BoW) para vetores de características visuais, capazes de expor a semântica latente entre itens (tomadas) similares. Quando aplicado a segmentação em cenas, cada tomada é representada por um histograma de ocorrência das "palavras visuais" que formam o dicionário visual, obtido por meio de um algoritmo de agrupamento dos vetores de características extraídos.

No trabalho específico de Chasanis, Kalogeratos e Likas (2009), um conjunto de quadros-chave são selecionadas para cada tomada, segmentadas manualmente, do vídeo. Para isso, os histogramas HSV 8:4:4 de cada quadro da tomada são agrupados por meio de um algoritmo de k-means, no qual os medóides de cada grupo (o vetor de característica mais próximo ao centro matemático do grupo) são selecionados como quadro-chaves.

Para cada quadro-chave, são extraídos características locais de imagem, como o SIFT e o *Contrast Context Histogram* (CCH) (HUANG; CHEN; CHUNG, 2008). Todos os vetores de características extraídos são, então, agrupados usando o modelo BoW, gerando um dicionário visual de tamanho k predefinido. Com tal dicionário, procede-se com o cálculo dos histogramas de ocorrências de palavras visuais para cada tomada. O resultado é um histograma VH_i, de tamanho k, que representa a i-tomada do vídeo de entrada. Segundo os autores, os histogramas gerados pela BoVW desconsideram qualquer informação presente nas tomadas adjacentes. Assim, um processo de suavização dos histogramas é realizado de acordo com a Equação 3.5 a seguir.

$$SH_t = \sum_{n=-\infty}^{\infty} VH_{t-n} \cdot K_\sigma(t-n) \tag{3.5}$$

Onde SH_t é o histograma suavizado da tomada t e K_σ é uma máscara ou *kernel* Gaussiano com média zero e desvio padrão σ.

Finalmente, a segmentação em cenas é realizada por meio do cálculo da distância euclidiana entre os histogramas suavizados *SH* entre tomadas adjacentes. Caso o valor da dissimilaridade seja um máximo local, a transição de cena é identificada.

Para avaliar a proposta, os autores utilizaram diferentes tamanhos de histogramas BoVW, contendo 10, 20, 50, 100, 200 e 500 palavras visuais, tanto com vetores de características SIFT e CCH. Usando as medidas de Precisão e Abrangência, três trechos dos filmes **A Beautiful Mind**, **Sex and the City** e **Gone in 60 seconds** foram avaliados, totalizando 3 horas e 6 minutos de vídeo. Assim como nas abordagens previamente descritas, tanto a base confiável como a sua implementação não estão disponíveis publicamente.

Tal abordagem também foi utilizada como base para o desenvolvimento de técnicas posteriores. Lopes, Trojahn e Goularte (2014), por exemplo, propuseram uma extensão a técnica BoVW ao usar uma abordagem chamada *Bag of Features* (BoF) áudio-visual. Na técnica desenvolvida, características aurais (MFCC) e visuais (SIFT) são extraídas de cada tomada e, usando diferentes BoF para cada modalidade, são construídos histogramas de ocorrência de palavras aurais e visuais. Tais histogramas, de maneira semelhante a técnica de Chasanis, Likas e Galatsanos (2009), são usadas para gerar segmentações em cenas unimodais que, após um processo de fusão tardia, são unidas em uma segmentação final. Tal abordagem multimodal, segundo os autores, resultou em melhoras significativas na segmentação quando comparado a abordagem baseada puramente em características visuais.

3.2 Técnicas em estado da arte

As abordagens seminais discutidas anteriormente costumam explorar apenas uma única modalidade, geralmente visual, para obter a segmentação em cenas. O desenvolvimento de técnicas mais recentes derivadas das abordagens iniciais aumentou a eficácia da segmentação em cenas em diferentes vídeos. Como uma evolução natural pesquisadores voltaram seu foco para o desenvolvimento de abordagens multimodais. Estas, em grande parte, se inspiraram nas abordagens unimodais previamente discutidas, como a análise de grafos (YEUNG; YEO; LIU, 1998), coerência entre tomadas (RASHEED; SHAH, 2003) ou histogramas de ocorrências de palavras visuais (CHASANIS; KALOGERATOS; LIKAS, 2009), para estabelecer meios de representação e comparação entre tomadas.

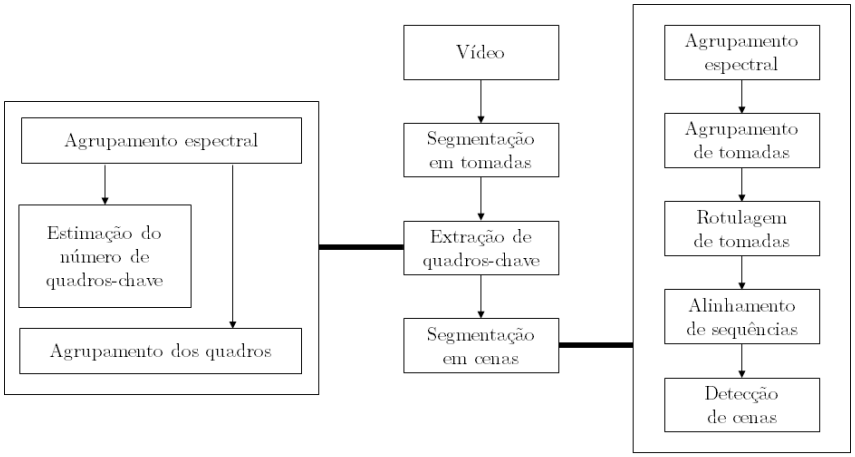
O estado da arte pode ser dividido em técnicas unimodais minuciosamente ajustadas para a tarefa de segmentação e técnicas multimodais. Essas últimas ainda possuem margem para melhorias significativas em eficácia e são, portanto, tema de pesquisas recentes. Elas podem ser divididas em técnicas baseadas e não baseadas em Aprendizagem Profunda. Nesta seção serão apresentadas técnicas pertencentes a essas classes (unimodais, multimodais não baseadas em Aprendizagem Profunda e multimodais baseadas em Aprendizagem Profunda), com o objetivo de realizar futuras comparações com a técnica proposta. Os critérios para a seleção dessas técnicas

foram: possuir eficácia comparável às melhores técnicas reportadas na literatura; terem resultados reportados sobre uma base pública e com métricas adequadas. Dentre as abordagens reportadas na literatura que atendem aos critérios supracitados, estão as técnicas: unimodal baseadas no alinhamento de sequências de Chasanis, Likas e Galatsanos (2009); o grafo de transição de cenas multimodal de Sidiropoulos *et al.* (2011); a técnica multimodal usando a Aprendizagem Profunda por meio de uma rede siamesa (BARALDI; GRANA; CUCCHIARA, 2015a). Tais técnicas serão, respectivamente, descritas nas subseções a seguir.

3.2.1 Técnica baseada no alinhamento de sequências

Inspirado na técnica de agrupamento de tomadas por meio de sua coerência visual proposta por Rasheed e Shah (2003), descrita previamente na Subseção 3.1.2, Chasanis, Likas e Galatsanos (2009) propuseram uma técnica de segmentação em cenas unimodal baseada no alinhamento de sequências de rótulos atribuídos às tomadas do vídeo. O diagrama apresentado na Figura 15 ilustra a técnica desenvolvida.

Figura 15 – Diagrama da técnica unimodal de segmentação de cenas, baseada no alinhamento de sequências, proposta por Chasanis, Likas e Galatsanos (2009).



Fonte: Adaptada de Chasanis, Likas e Galatsanos (2009).

A abordagem se baseia na segmentação em tomadas do vídeo, realizada por meio de comparação de histogramas de cor. Posteriormente, para cada tomada do vídeo, são extraídos um número variável (estimado automaticamente) de quadros-chave baseado na distância euclidiana entre histogramas de cor HSV, calculados por meio de um algoritmo de agrupamento espectral (NG; JORDAN; WEISS, 2001). Tal abordagem também foi adotada em outros trabalhos relacionados, como os desenvolvidos por Lopes, Trojahn e Goularte (2014) e Chasanis, Kalogeratos e Likas (2009), descritos previamente.

Com os quadros-chave selecionados é calculado a similaridade visual, dada pelo maior valor de interseção entre os histogramas de cor HSV entre as tomadas duas-a-duas. Seja K_i e K_j o conjunto de quadros-chave das tomadas i e j, H_a e H_b os histogramas normalizados de cor

HSV dos quadros-chave a e b, então a similaridade visual (do inglês *Visual Similarity* - *VisSim*) entre duas tomadas i e j, e a similaridade de cor (do inglês *Color Similarity* - *ColSim*) entre quadros-chave a e b são dadas por:

$$VisSim(i,j) = \max_{p \in K_i, q \in K_j} ColSim(p,q) \tag{3.6}$$

$$ColSim(a,b) = \sum_{h \in bins} \min(H_a(h), H_b(h)) \tag{3.7}$$

Após calculadas as similaridades entre cada tomada (*VisSim*), é calculada uma matriz de similaridade de tamanho *NxN*, considerando um vídeo com N tomadas, usado pelo algoritmo de agrupamento espectral. A cada grupo é atribuído um rótulo único, que identificará todas as suas respectivas tomadas. Finalmente, dado o rótulo de cada tomada do vídeo, é utilizado o algoritmo de alinhamento de sequências de Needleman-Wunsch (NW)(NEEDLEMAN; WUNSCH, 1970), aplicado na área de bioinformática para o alinhamento de sequências de proteínas ou nucleotídeos, detectando os pontos de dissimilaridade entre sequências. Tais pontos de dissimilaridade são comparados a um limiar, correspondente a 80% da maior dissimilaridade global encontrada, e, caso sejam menores que tal limiar, são considerados como transições de cenas.

Em termos de avaliação, os autores utilizaram uma base de dados customizada. Tal base, formada por 5051 tomadas, 177 cenas e cerca de 5 horas de duração, é composta de dez vídeos no domínio de séries de TV e filmes nos gêneros de comédia, drama e ação. Os resultados obtidos, baseados nas métricas de Precisão e Abrangência, foram comparados com as técnicas unimodais propostas por Yeung, Yeo e Liu (1998) e Rasheed e Shah (2005). A Tabela 3 apresenta os resultados médios reportados por Chasanis, Likas e Galatsanos (2009).

Tabela 3 – F_{PR} médio e desvio padrão reportados por Chasanis, Likas e Galatsanos (2009) em uma base de vídeos customizada.

	F_{PR}	Desvio Padrão
Chasanis, Likas e Galatsanos (2009)	84.26%	5.85
Yeung, Yeo e Liu (1998)	70.74%	3.92
Rasheed e Shah (2005)	71.29%	4.45

Infelizmente, a base de vídeos não está disponível publicamente, o que inviabiliza o uso da mesma como modo de comparação nas avaliações realizadas neste trabalho. Contudo, a comparação com este trabalho foi possível por meio de resultados reportados por outros autores, o que será discutido posteriormente.

3.2.2 Técnica baseada no grafo de transição de cena multimodal

Sidiropoulos *et al.* (2011), por sua vez, apresentam uma extensão do trabalho de Yeung, Yeo e Liu (1998) acrescentando ao STG análise multimodal baseada em quatro diferentes características: histogramas de cor e de classificadores de áudio (baixo nível); além de classificadores

de imagem e de eventos aurais (alto nível). A técnica consiste em obter segmentações mais representativas, por meio da multimodalidade, além de se tornar independente de configurações dos parâmetros da STG ao gerar múltiplos grafos com diferentes conjuntos de parâmetros. O Algoritmo 1 descreve o algoritmo de criação do STG dado um conjunto de tomadas $S = s_1, ..., s_n$, seus respectivos histogramas ou vetores de características $H = h_1, ..., h_n$, o tamanho de janela ω e um limiar de agrupamento δ.

Após construídos os diversos grafos para cada uma das características utilizadas, calcula-se, para cada tomada do vídeo, um grau de confiança por meio do número de grafos que a classificam como uma transição de cena, dividida pelo número total de grafos daquela característica. Em outras palavras, é gerado um valor para determinar em quantos grafos tal tomada foi considerada como transição de cena. Após calculados todos os graus de confiança para cada característica (p^y), considerando ω_y o peso associado a classe de característica y, sendo $\sum \omega_y = 1$, a fusão tardia é realizada seguindo a Equação 3.8.

$$p_i = \sum_y \omega_y \cdot p_i^y \tag{3.8}$$

Finalmente, as transições de cenas Γ são detectadas por meio da comparação entre as tomadas sequenciais x_i e x_{i+1} e um limiar T predefinido, tal que:

$$\Gamma = \{(x_i, x_{i+1}) | p_i > T\} \tag{3.9}$$

Em termos de avaliação, os autores utilizaram duas bases de vídeos cujas transições de cenas foram geradas manualmente: uma base contendo quinze documentários do *Netherlands Institute for Sound and Vision*[1], totalizando 8 horas e 33 minutos de vídeo, 3459 tomadas e 525 cenas; além de uma base contendo seis filmes, não especificados, totalizando 10 horas e 43 minutos de vídeo, 6665 tomadas e 357 cenas. Os resultados médios obtidos, reportados usando as métricas de Cobertura e Transbordamento descritas anteriormente na Subseção 2.6.2, foram ainda comparados com outros trabalhos presentes na literatura, incluindo a técnica baseada em alinhamento de sequências de Chasanis, Likas e Galatsanos (2009) descrita na Subseção 3.2.1, além das técnicas propostas por Nitanda, Haseyama e Kitajima (2005) e Wilson e Divakaran (2009). A Tabela 4 descreve os resultados reportados por Sidiropoulos *et al.* (2011).

Tabela 4 – Resultados de F_{CO} médio usando duas bases de vídeos diferentes, reportados por Sidiropoulos *et al.* (2011).

	Documentários	Filmes
Sidiropoulos *et al.* (2011)	87.67%	84.91%
Nitanda, Haseyama e Kitajima (2005)	80.06%	75.41%
Chasanis, Likas e Galatsanos (2009)	73.30%	79.97%
Wilson e Divakaran (2009)	80.67%	79.16%

[1] <http://instituut.beeldengeluid.nl>

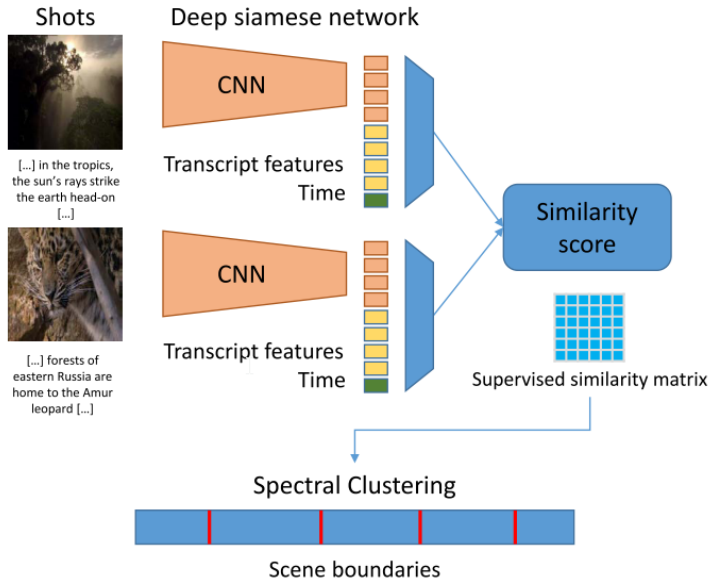
Infelizmente, as segmentações em cenas das duas bases utilizadas, tanto a de documentários como a de filmes, não estão disponíveis publicamente, impossibilitando o uso das mesmas nas avaliações realizadas neste trabalho. Contudo, a comparação com este trabalho foi possível por meio de resultados reportados por outros autores, o quê será discutido posteriormente.

3.2.3 Técnica baseada em rede neural siamesa

Dentre as abordagens de segmentação em cenas sem domínio específico, a técnica baseada em uma rede neural siamesa profunda (do inglês *Siamese Deep Network* - SDN), proposta por Baraldi, Grana e Cucchiara (2015a), é uma das primeiras abordagens a utilizar a abordagem de Aprendizagem Profunda e que reporta resultados em estado da arte em relação a métrica da média harmônica F-measure entre Cobertura e Transbordamento (F_{CO}).

A técnica proposta, ilustrada na Figura 16, consiste em uma rede siamesa, cujos pesos e parâmetros são idênticos, responsáveis por receber características de duas tomadas diferentes e calcular um valor de similaridade entre as mesmas. Por meio da análise tomada-a-tomada, resultando em $\frac{n^2-n}{2}$ execuções da rede, uma matriz de similaridade entre todas as tomadas é obtida. Tal matriz passa, então, por um processo de agrupamento espectral, de maneira a agrupar as tomadas e identifica as transições de cena.

Figura 16 – Ilustração da técnica de segmentação de cenas, baseada na análise da similaridade entre tomadas, por meio de uma rede convolucional siamesa.



Fonte: Baraldi, Grana e Cucchiara (2015a).

A CNN utiliza uma arquitetura baseada na AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), usada como referência no *framework* Caffe (JIA *et al.*, 2014), formada por camadas de convolução, ReLU e *max pooling*, treinada em duas bases de dados de imagens bem conhecidas, com o objetivo de detectar objetos e localização (exterior/interior) da imagem. O re-

sultado da extração de tais características visuais para cada quadro-chave da tomada, selecionado o quadro mediano de cada tomada, é um vetor único contendo 1183 valores.

Além das características visuais, a técnica utiliza ainda histogramas de ocorrências de palavras textuais. Para isso, cada palavra presente na legenda é representada por um vetor de característica Word2Vec (Mikolov *et al.*, 2013), no qual todos os vetores extraídos são então agrupados usando o algoritmo k-means baseado na distância do cosseno. Seguindo o modelo BoW, os vetores de características são usados para criar um "dicionário", usado para o cálculo de um histograma de ocorrências para cada tomada. Segundo os autores, devido a baixa duração de cada tomada, uma janela deslizante é adotada, chamada de janela contextual, na qual as palavras contidas entre cada janela são usadas em todas as legendas da mesma janela. O tamanho de cada histograma é de 200 valores, enquanto que o tamanho da janela contextual adotada é de 20 segundos de duração. Por fim, o índice do quadro mediano, também utilizado para a extração das características visuais, é utilizado para representar o "tempo" na rede proposta.

Após extraídas, as características são então concatenadas em um vetor único para cada ramo da rede siamesa. Tais vetores são analisados por uma rede totalmente conectada que estima a distância entre as tomadas, resultando na matriz de similaridade previamente mencionada que, após um processo de agrupamento temporal, é utilizada para a segmentação em cenas.

Além da técnica, os autores propõem uma base de dados disponibilizada publicamente chamada de BBC *Dataset*, descrita anteriormente. A eficácia da técnica foi realizada comparando-a com duas outras técnicas consideradas em estado da arte: a técnica unimodal baseada em alinhamento de sequências (Subseção 3.2.1) e a técnica multimodal baseada no grafo de transição de cenas (Tabela 5). Os resultados usaram as métricas de Cobertura e Transbordamento, superando as referidas técnicas conforme descrito na Tabela 5.

Tabela 5 – Resultados médios reportados por Baraldi, Grana e Cucchiara (2015a) ao realizar a segmentação em cenas sobre a base de vídeos BBC *Dataset*, comparando-a com as técnicas baseada em grafo de transição de cenas de Sidiropoulos *et al.* (2011) e de alinhamento de sequências de Chasanis, Likas e Galatsanos (2009).

	F_{CO}
Baraldi, Grana e Cucchiara (2015a)	62%
Chasanis, Likas e Galatsanos (2009)	45%
Sidiropoulos *et al.* (2011)	54%

É importante ressalvar que os resultados obtidos pelas técnicas propostas por Chasanis, Likas e Galatsanos (2009) e Sidiropoulos *et al.* (2011) são inferiores aos obtidos originalmente por seus autores. Uma possível explicação para tanto é a maior complexidade da BBC *Dataset* quando comparada com as bases de vídeos originalmente adotadas. Nos trabalhos originais, os autores avaliaram suas técnicas sobre bases de dados de trechos de filmes e séries de TV, que podem favorecer em demasia suas abordagens. Já a avaliação realizada por Baraldi, Grana e Cucchiara (2015a) consiste em avaliar o vídeo completo, o que tende a evidenciar limitações

das técnicas devido a presença de cenas complexas e com comportamento diferenciado em uma mesmo vídeo.

Devido a disponibilidade da BBC *Dataset*, tanto de seus vídeos como as respectivas segmentações manuais confiáveis, além do próprio algoritmo de comparação das segmentações obtidas, os resultados reportados por Baraldi, Grana e Cucchiara (2015a) foram utilizados como modo de comparação com os resultados obtidos pelas redes desenvolvidas, conforme avaliação apresentada no Capítulo 5. É importante ressalvar que apenas o valor F_{CO} foi reportado, não apresentando os valores individuais das métricas de Cobertura C ou Transbordamento O.

3.3 Discussões sobre o capítulo

Neste capítulo foram descritas diversas técnicas reportadas na literatura acerca da segmentação de vídeo em cenas sem domínio específico, foco desta proposta. Foram descritas as técnicas mais relacionadas ao presente trabalho, tanto seminais como em estado da arte, usando abordagens unimodais ou multimodais, estas baseadas ou não em Aprendizagem Profunda.

As técnicas descritas na Seção 3.1 retratam abordagens seminais para a tarefa da segmentação em cenas, que apresentam limitações significativas ou cujos resultados alcançados foram superados por técnicas posteriores. Dentre as técnicas seminais presentes na literatura, é perceptível a predominância da abordagem unimodal, particularmente de técnicas que procuram medir a similaridade visual tomada-a-tomada. Rasheed e Shah (2003) e Trojahn e Goularte (2013), por exemplo, utilizam histogramas comparados em uma janela deslizante para determinar a similaridade de uma dada tomada frente as anteriores. Chasanis, Kalogeratos e Likas (2009), por sua vez, utilizaram o modelo BoW aplicado a características locais de imagens como o SIFT e CCH para estimar um histograma de ocorrências de palavras visuais, usado para representar a tomada e determinar a similaridade com outras tomadas, detectando transições de cenas. Lopes, Trojahn e Goularte (2014), posteriormente, propuseram uma extensão multimodal de tal técnica, utilizando ainda vetores MFCC (aurais) e SIFT (visuais).

Já na Seção 3.2 foram apresentadas técnicas cujos resultados alcançados podem ser considerados em estado da arte para suas respectivas abordagens. Nesse sentido, das abordagens não baseadas em Aprendizagem Profunda, foram descritas uma unimodal, baseada no alinhamento de sequências de rótulos de tomadas (CHASANIS; LIKAS; GALATSANOS, 2009), além da abordagem multimodal do grafo de transição de cenas (SIDIROPOULOS *et al.*, 2011). Posteriormente, uma técnica multimodal em estado da arte usando a Aprendizagem Profunda foi descrita, baseada em redes neurais siamesas (BARALDI; GRANA; CUCCHIARA, 2015a). Tais técnicas reportam resultados utilizando uma base de vídeos adequada, publicamente disponível, para a tarefa de segmentação em cenas. Por isso, tais técnicas foram utilizadas para fins de comparação com o método proposto neste trabalho.

As técnicas reportadas, tanto seminais como em estado da arte, mostram a tendência

recente da adoção da multimodalidade, haja visto as limitações na detecção do contexto e/ou do assunto usando apenas uma única modalidade. É perceptível, também, a adoção de abordagens baseadas em Aprendizagem Profunda nos últimos anos devido, principalmente, a seu grande sucesso em áreas correlatas, como a classificação de imagens. É possível notar, ainda, uma tendência pela adoção de características de maior nível semântico, tais como as características convolucionais e histogramas de ocorrência de palavras adotados por Sidiropoulos *et al.* (2011) e Baraldi, Grana e Cucchiara (2017), devido ao seu potencial de melhor detectar o contexto, obtendo melhores e menos limitadas segmentações em cenas.

Uma lacuna de pesquisa perceptível nos trabalhos relacionados para a segmentação em cenas é a ausência de consenso de como representar o *tempo*. Como uma cena é formada por uma sequência temporal de tomadas, é de interesse dos pesquisadores criarem algoritmos que se adaptem a mais ampla gama de combinações temporais possíveis, tais como cenas de longa ou curtíssima duração. Yeung, Yeo e Liu (1998) e Sidiropoulos *et al.* (2011), por exemplo, utilizam o tempo para determinar as arrestas entre os grupos (vértices) de tomadas, cujas arrestas de corte são usadas para detectar transições de cenas. Rasheed e Shah (2003) e Trojahn e Goularte (2013), utilizam a informação temporal indiretamente, calculando a similaridade de uma tomada frente a outras usando uma janela deslizante de tamanho predefinido. Chasanis, Kalogeratos e Likas (2009) e Lopes, Trojahn e Goularte (2014), percebendo a importância das informações temporais, propõem a suavização dos histogramas dada as tomadas adjacentes. No caso particular da técnica baseadas em Aprendizagem Profunda usando redes siamesas (BARALDI; GRANA; CUCCHIARA, 2015a), o tempo é representado como o número do quadro mediano da tomada. Nesse sentido, o modelo que compõe o método de segmentação em cenas proposto conta com uma rede neural recorrente especialmente eficaz em identificar padrões temporais presente em tomadas adjacentes.

Dentre as técnicas multimodais que não usam a Aprendizagem Profunda, é perceptível a predominância da abordagem da fusão tardia, como as técnicas reportadas por Sidiropoulos *et al.* (2011) e Lopes, Trojahn e Goularte (2014), principalmente devido a dificuldades de modelar e manipular informações ou características provenientes de modalidades diferentes. Por outro lado, as técnicas baseadas na Aprendizagem Profunda propõem arquiteturas baseadas na fusão antecipada (BARALDI; GRANA; CUCCHIARA, 2015a; BARALDI; GRANA; CUCCHIARA, 2017), evidenciando a alta flexibilidade que uma rede profunda possui de lidar com características de diferentes naturezas. Não há, porém, qualquer trabalho encontrado que mencione ou justifique tal escolha. A fusão tardia permite uma maior flexibilidade na análise de cada modalidade, uma característica desejável e até mesmo fundamental em diversos cenários como processamento de vídeo distribuído. Nesse sentido, o desenvolvimento de um método de fusão flexível, que permite facilmente a criação de uma abordagem baseada na fusão antecipada ou tardia, poderia auxiliar a diferentes tarefas multimodais relacionados a vídeo, tal como a segmentação em cenas. Assim, o modelo proposto neste trabalho pode ser instanciado usando tanto a fusão antecipada como tardia sem necessitar modificações das etapas de Pré-processamento ou Segmentação em si.

CAPÍTULO

4

DESCRIÇÃO DO MÉTODO PROPOSTO

Este capítulo apresenta o método de segmentação em cenas proposto. O método é formado pela modelagem do problema da segmentação temporal de vídeo em cenas como um problema de classificação de tomadas, conforme descrito na Seção 4.1, além de um modelo formado pela associação de redes neurais convolucionais e redes neurais recorrentes descrito na Seção 4.2.

As redes neurais propostas oferecem um alto grau de flexibilidade quanto a abordagem de fusão multimodal, podendo ser aplicadas em duas arquiteturas baseadas na fusão antecipada ou tardia, conforme detalhado na Seção 4.3. Adicionalmente, as redes desenvolvidas suportam uma ampla gama de diferentes características de entrada, sendo, assim, instanciadas por meio de quatro características de três modalidades diferentes, detalhado na Seção 4.4. Após a classificação da tomada de entrada por uma rede neural proposta, faz-se necessário realizar um processo de segmentação em cenas com o intuito de detectar as tomadas de transição e obter a segmentação em si, conforme descrito na Seção 4.5. Por fim, uma discussão sobre este capítulo é apresentada na Seção 4.6.

4.1 Modelagem do problema

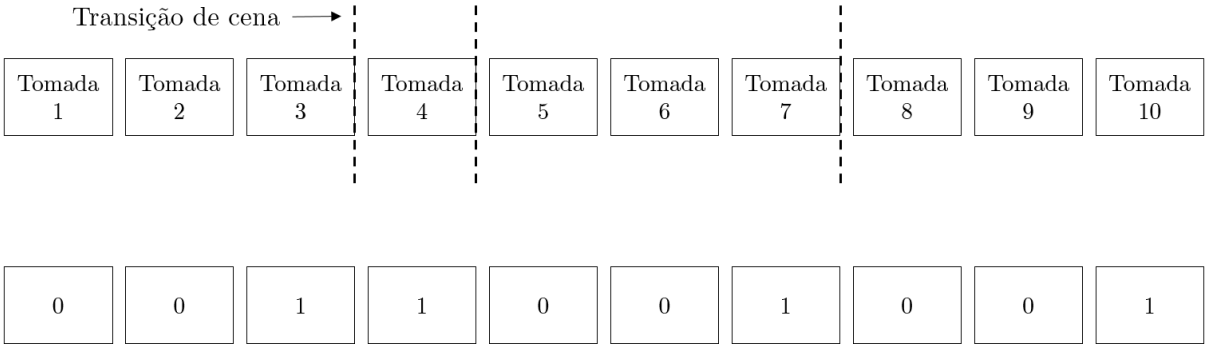
O problema da segmentação de vídeo em cenas é o de identificar o momento no qual há uma transição de cena. Neste trabalho, foi adotada a definição na qual uma “cena” é uma sequência de tomadas semanticamente relacionadas, conforme discutido na Seção 2.1. Assim, a transição de cenas ocorre concomitantemente à transição entre duas tomadas. Com isso, a segmentação em cenas pode ser modelada como um problema de classificação binária de tomadas: as que são de transição e as que não são. Nessa modelagem, as tomadas de transição são aquelas que encerram a respectiva cena, enquanto que tomadas de não-transição estão localizadas entre tomadas de transição diferentes.

Seja um vídeo V composto por n tomadas, T_i uma tomada do vídeo V e T_{trans} o conjunto de tomadas de transição do vídeo V, então cada tomada T_i é classificada como 1 caso seja de transição ou 0 caso contrário. Ou seja:

$$T_i = \begin{cases} 1, & \text{se } T_i \in T_{trans} \text{ou } i = n \\ 0, & \text{caso contrário} \end{cases} \tag{4.1}$$

Para ilustrar tal modelagem, considere um vídeo composto de dez tomadas e de três transições de cenas, conforme ilustrado na Figura 17. Assim, a sequência binária de classificação entre tomadas de transição ou não seria 0011001001.

Figura 17 – Exemplo da modelagem do problema de segmentação em cenas como um problema de classificação de tomadas. Os retângulos representam as tomadas e as linhas verticais denotam as transições de cenas.



Fonte: Elaborada pelo autor.

É importante ressalvar que a última tomada do vídeo é classificada como de transição para fins de normalização, haja visto que a mesma encerra a última cena do vídeo. A utilização de tal modelagem resulta em alguns benefícios para a segmentação em cenas:

- Vídeos de qualquer tamanho e com qualquer número de cenas podem ser representados, inclusive aqueles que não possuem qualquer transição de cenas.
- A representação suporta cenas de quaisquer tamanhos, inclusive as compostas por apenas uma única tomada.
- A modelagem em si não está atrelada a qualquer domínio ou definição de cena em particular, sendo, portanto, aplicável em diferentes situações.

Uma consequência indireta ao adotar tal modelagem para representar o problema da segmentação em cenas é o potencial ganho de desempenho, já que a classificação de todas as tomadas de um vídeo pode ser obtida com apenas uma única execução da rede. Tal característica contrasta com técnicas relacionadas (como a proposta por Baraldi, Grana e Cucchiara (2015a)) na qual é necessário executar o processo de classificação da rede proposta diversas vezes

para calcular a matriz de similaridade requerida pelo algoritmo de agrupamento que obterá a segmentação em cenas.

4.2 Modelo proposto

O modelo proposto neste trabalho é formado da associação de redes neurais convolucionais (CNN) e redes neurais recorrentes (RNN) para realizar a classificação de uma dada tomada de entrada, independentemente da modalidade ou característica de entrada, seguindo a modelagem do problema apresentado anteriormente. Com tal modelo, os vetores de características extraídos das diferentes modalidades são analisados pela CNN, resultando em um vetor de característica único para cada tomada de entrada. A RNN é responsável por, usando tais vetores de características únicos, identificar o correlacionamento temporal e realizar a classificação de cada tomada de entrada.

A utilização de CNNs e RNNs resulta em um modelo com diversos benefícios para a tarefa de segmentação em cenas. As CNN são altamente eficazes em detectar e extrair padrões significativos dos dados ou características de entrada, oferecendo uma boa representação para cada tomada do vídeo que, inclusive, auxilia a identificação do correlacionamento temporal na rede recorrente posterior. Devido a suas particularidades, as CNNs são capazes de analisar características provenientes de diferentes modalidades, como aural e visual, tornando a técnica de segmentação, a RNN e o modelo em si flexíveis quanto as características de entrada. Adicionalmente, com a alteração de um pequeno conjunto de parâmetros como número de camadas convolucionais e número de neurônios de entrada, a CNN oferece suporte a uma ampla gama de características de diferentes modalidades. A CNN empregada no modelo proposto é detalhado na Subseção 4.2.1.

A RNN, por sua vez, é responsável por analisar a representação da tomada gerada anteriormente pela CNN para classificá-la de acordo com a modelagem proposta, potencialmente identificando padrões que possam indicar a mudança de assunto e, consequentemente, de cenas do vídeo. Por se tratar de uma rede recorrente, seu principal benefício para a segmentação em cenas é a capacidade de, ao classificar uma tomada, efetivamente lembrar de várias tomadas anteriores, resultando em uma análise apurada do correlacionamento presente na representação utilizada como entrada. Graças a extração de características realizada pela CNN, a RNN independe das características de entrada, sendo capaz de analisar e classificar dados provenientes de diferentes modalidades em qualquer combinação. Tal capacidade também resulta na flexibilidade do modelo proposto quanto ao processo de fusão multimodal, suportando tanto a fusão antecipada como a fusão tardia por meio de pequenas alterações na arquitetura da rede. A RNN empregada no modelo proposto é detalhado na Subseção 4.2.2.

4.2.1 Rede convolucional

A CNN desenvolvida no modelo proposto neste trabalho é responsável por analisar as características de entrada de cada modalidade individual e destacar padrões significativos para a tarefa de segmentação em cenas. Objetiva-se, também, a reduzir o volume de informações de entrada, padronizando cada modalidade e gerando uma representação única e adequada para sua posterior classificação em uma rede neural recorrente.

A CNN desenvolvida se baseia nas arquiteturas de redes propostas da VGGNet (SIMONYAN; ZISSERMAN, 2014) e AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), redes convolucionais criadas para a classificação de imagens. É importante salientar que embora baseada nelas, a CNN proposta possui propriedades únicas que a distingue das referidas redes. Tanto a VGGNet como a AlexNet utilizam como entrada os valores dos pixels RGB de uma imagem, aplicando a *convolução espacial* para a extração de características significativas para sua posterior classificação. Já na CNN desenvolvida, cada modalidade da tomada é representada por um conjunto de características unidimensional, sendo aplicado então uma operação *convolução temporal* para a extração de características da entrada.

Conforme previamente descrito na Subseção 2.4.1, uma CNN é composta de combinações de camadas de convolução, camadas não-lineares e de subamostragem. Com isso, dado como entrada de cada rede o conjunto de características de baixo-nível de cada modalidade específica, é possível obter uma representação dimensionalmente reduzida para cada tomada do vídeo. No modelo proposto, a CNN utiliza a seguinte configuração:

- Convolução temporal (1D) de amostras duas-a-duas.
- *Max pooling* temporal (1D).
- ReLU, definida como $f(x) = \max(0, x)$ (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), como a camada não-linear da convolução.

A camada convolucional é responsável por destacar os padrões mais relevantes obtendo, potencialmente, uma abstração de maior nível dos dados de entrada. Foi adotado um tamanho de máscara ou *kernel* proporcionalmente pequeno ao número de vetores de entrada, prática esta também empregada em trabalhos relacionados (SIMONYAN; ZISSERMAN, 2014; HE *et al.*, 2016). É importante destacar ainda que os valores específicos de cada máscara ou *kernel* de cada camada convolucional são determinados automaticamente pela própria rede convolucional em tempo de treinamento.

Já a camada de *max pooling* visa realizar a subamostragem temporal da entrada, resultando em redução do volume de dados. Tal camada é responsável, ainda, por remover a potencial influência que a duplicação de características de entrada, processo adotado neste trabalho quando o número de vetores de características extraídos em uma tomada é menor que o esperado, poderia

haver sobre a rede como um todo. Assim, tal camada provê uma maior flexibilidade quanto a etapa de pré-processamento, na qual as características do vídeo são extraídas, suportando assim uma maior gama de possíveis características de entrada.

Por fim, a camada não-linear ReLU, quando associada a uma camada convolucional, é capaz de remover a influência de padrões ruidosos ou de menor relevância, resultando em abstrações de maior nível semântico. A ReLU, em particular, foi adotada por obter resultados satisfatórios em vários trabalhos relacionados, além de sua alta velocidade de convergência (XU *et al.*, 2015), resistência a sobre-especialização e grande popularidade (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; SIMONYAN; ZISSERMAN, 2014; HE *et al.*, 2016; HE *et al.*, 2015; NAIR; HINTON, 2010).

Com o uso sequencial de múltiplos conjuntos de camadas seguindo a arquitetura previamente especificada, as características são analisadas até obter um único vetor de característica que será provida a uma RNN para posterior classificação. É importante destacar que, no modelo proposto, a aplicação da CNN para a extração de padrões significativos e obtenção de uma representação para a tomada é opcional, podendo uma determinada característica ou modalidade ser utilizada diretamente pela RNN, caso a característica seja composta de um único vetor de característica de tamanho fixo. Exemplos de tais casos são histogramas de ocorrência de palavras textuais (*Bag of Words*), histogramas de cor médio da tomada, entre outros.

4.2.2 Rede recorrente

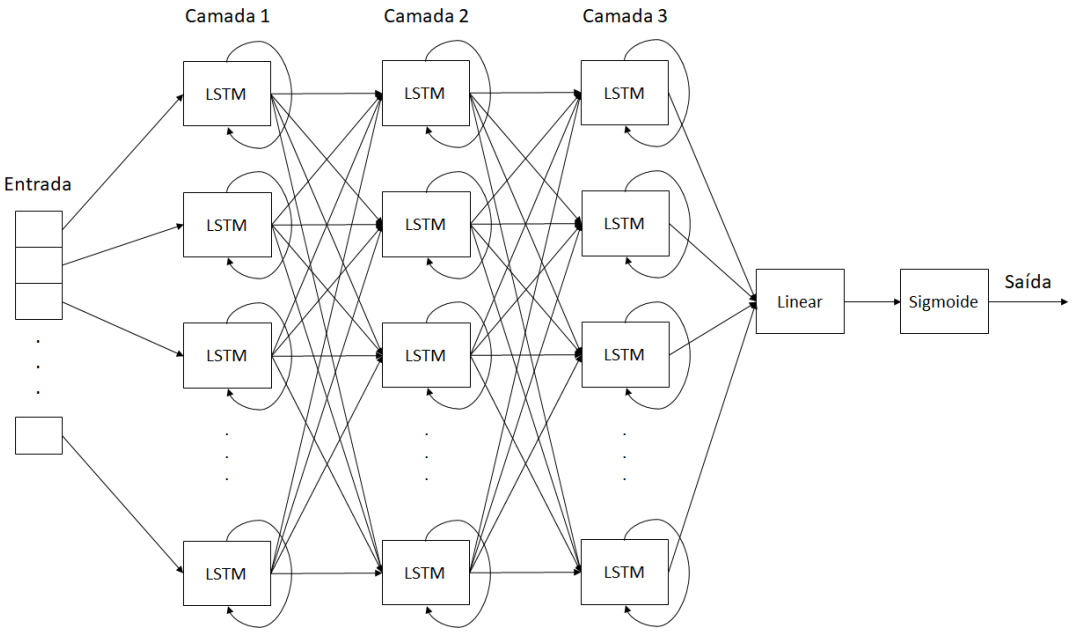
No modelo proposto, a RNN tem por objetivo identificar padrões significativos da representação das tomadas de entrada, classificando-as com o intuito de detectar as transições de cenas, conforme a modelagem proposta. Para auxiliar a classificação, a RNN busca ainda encontrar o correlacionamento temporal entre tomadas adjacentes, que podem indicar o relacionamento semântico entre tomadas de uma mesma cena.

Para permitir a utilização de uma ampla gama de características de entrada, além do suporte a diferentes abordagens de fusão multimodal, a RNN foi desenvolvida de maneira flexível e altamente adaptável a diversas situações. Estruturalmente, a RNN proposta neste trabalho é composta de três camadas de unidades LSTM. Tal arquitetura é similar ao adotada em CNNs como a VGGNet e AlexNet, embora suas camadas totalmente conectadas, utilizadas para classificação das imagens de entrada, sejam compostas de unidades ou neurônios simples com funções de ativação sigmoidais ou tangenciais.

Conforme previamente mencionado, cada unidade LSTM utilizada é baseada no trabalho de Jozefowicz, Zaremba e Sutskever (2015), cuja formulação é descrita na Subseção 2.4.2. A Figura 18 ilustra a rede neural recorrente adotada.

Na Figura 18, cada valor do vetor de característica de entrada é associado a uma unidade LSTM da primeira camada da rede desenvolvida. Após o processamento dos dados de entrada

Figura 18 – Ilustração da rede neural recorrente proposta, formada por três camadas de unidades LSTM, seguidas de uma camada linear e a operação de sigmoide.



Fonte: Elaborada pelo autor.

por três camadas recorrentes, os dados são agrupados em um único valor por meio de uma função linear sigmoidal sendo seu valor normalizado no intervalo aberto entre zero e um por uma função sigmoide, resultando no valor de classificação desejado para a tomada de entrada, de acordo com a modelagem proposta.

Na RNN proposta, dois parâmetros fundamentais foram avaliados: o número de camadas recorrentes e o número de elementos LSTM em cada camada, além do tamanho da memória de cada elemento LSTM. No primeiro caso, o número de camadas e de elementos LSTM em cada camada foi determinado heuristicamente utilizando diferentes configurações, encontrando-se o valor de 500 unidades LSTM por camada, distribuídas em três camadas. Não foi adotada nenhuma camada intermediária, como ReLU ou *Dropout*, entre as camadas recorrentes. Tal configuração apresenta uma boa relação custo/benefício, permitindo tanto a convergência no treinamento em tempo razoável, quanto uma eficácia equivalente a outras redes mais largas ou profundas (com maior número de elementos ou mais camadas).

Já o segundo parâmetro avaliado, a memória dos elementos LSTM, refere-se ao momento de corte do gradiente com base em um número predefinido de amostras anteriores em unidade LSTM, parâmetro este diretamente relacionado ao problema da explosão do gradiente do algoritmo de treinamento. Além de combater a explosão de gradiente, a memória de cada elemento LSTM indica o tamanho da janela de correlacionamento: dado um tamanho de memória N, para classificar uma tomada k serão utilizadas as informações contidas nas $k - N$ tomadas anteriores. Neste trabalho, diversos tamanhos de memória foram considerados, tais como 3, 5, 10, 50 ou infinitas tomadas, com pequeno impacto percebido na eficácia final da segmentação. Para fins de

padronização e considerando as características da tarefa de segmentação em cenas, nas quais tomadas mais próximas possuem maior influência do que tomadas muito distantes, o tamanho de memória de 5 tomadas foi escolhido empiricamente. Em testes de eficácia realizados durante o desenvolvimento, tal tamanho de memória não impactou negativamente os resultados de uma memória mais extensa, embora o uso de uma memória menor (principalmente memórias de uma ou duas tomadas) tenha resultado em perdas na ordem de 1% a 2% na eficácia média.

Após o processamento dos dados de entrada, a saída obtida pela RNN é então processada por uma função linear sigmoidal. Tal função é formada de uma camada que aplica uma operação linear sobre a entrada, agrupando os valores individuais em um único valor numérico. É aplicada a operação de sigmoide sobre tal valor, resultando, por fim, na classificação da tomada de entrada no intervalo aberto entre 0 e 1. A função linear sigmoidal $lsig(I)$ é apresentada na Equação 4.2, considerando I um vetor de entrada de tamanho n, W_k o peso do respectivo k-valor de entrada e b o valor de *bias* (opcional).

$$lsig(I) = \frac{1}{1 + e^{-(W_1 \cdot I_1 + W_2 \cdot I_2 + \ldots + W_n \cdot I_n + b)}} \tag{4.2}$$

A flexibilidade da RNN quanto a abordagem da fusão multimodal é resultado da capacidade de receber informações de qualquer natureza e de qualquer volume, necessitando apenas alterar o número de conexões na primeira camada da RNN. Assim, diferentes tipos de informações podem ser adequadamente classificados, incluindo representações unimodais obtidas pela CNN proposta, como qualquer combinação entre elas. Adicionalmente, como mencionado anteriormente, a RNN suporta como entrada, ainda, características que não tenham sido processadas por uma rede convolucional, desde que sejam formadas de apenas um único vetor de característica de tamanho fixo.

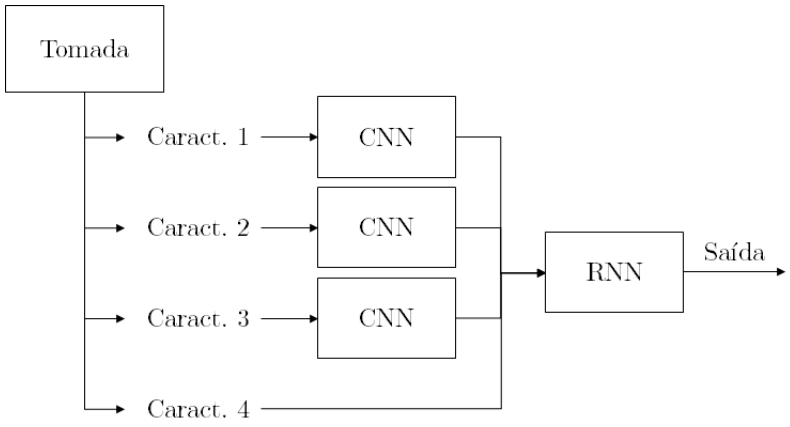
4.3 Arquitetura

De forma a instanciar o modelo proposto usando diferentes abordagens de fusão multimodal, duas arquiteturas são propostas, uma para a fusão antecipada e outra para a fusão tardia.

No caso da fusão antecipada, ao final da extração das características de cada modalidade na CNN, é criado um vetor de característica único multimodal, formado pela concatenação da representação unimodal obtida em cada CNN, que é então provido diretamente para a (única) RNN da arquitetura. Tal arquitetura é ilustrada na Figura 19.

Já a arquitetura do modelo de rede usando a fusão tardia se trata de um reconfiguração das conexões existentes, sem mudanças expressivas nas definições individuais de cada CNN ou RNN. Nessa arquitetura, ilustrada na Figura 20, para cada modalidade de entrada, a saída de sua correspondente CNN é provida diretamente para uma RNN individual que, sem ter acesso às demais modalidades, classificará a tomada de entrada de acordo com as informações contidas

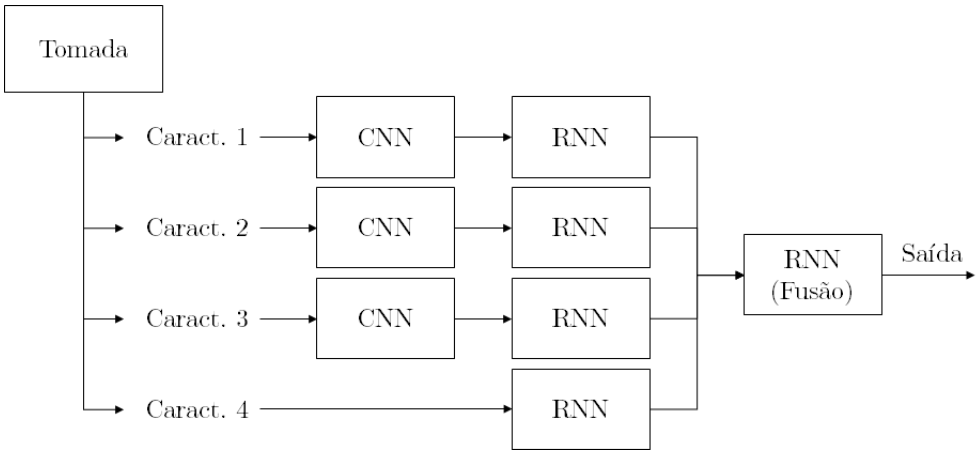
Figura 19 – Ilustração da arquitetura do modelo da rede neural proposta utilizando a fusão antecipada, considerando uma tomada de entrada e quatro diferentes características utilizadas. A saída é um valor numérico no intervalo entre 0 e 1. Note que a Característica 4 é provida diretamente à RNN sem necessitar da análise por uma CNN.



Fonte: Elaborada pelo autor.

em tal modalidade. Após a classificação unimodal, um processo de fusão é adotado para unir as decisões individuais com o objeto de obter a decisão ou segmentação final desejada. Note que a arquitetura é flexível no sentido de não restringir o como tal fusão é realizada, sendo possível tanto o uso de técnicas de Aprendizagem de Máquina (como uma nova RNN, conforme ilustrado na Figura 20), um algoritmo de ranqueamento preestabelecido ou outra abordagem qualquer.

Figura 20 – Ilustração da arquitetura do modelo da rede neural proposta utilizando a fusão tardia, considerando uma tomada de entrada e quatro diferentes características utilizadas. A saída é um valor numérico no intervalo entre 0 e 1. Note que a Característica 4 é provida diretamente à RNN sem necessitar da análise por uma CNN.



Fonte: Elaborada pelo autor.

Seguindo a arquitetura apresentada na Figura 20, a RNN utilizada para a fusão recebe como entrada um vetor contendo os quatro valores de classificação individuais no intervalo $[0.0, 1.0]$ para cada tomada, retornando a classificação final cujo valor também pertence ao

intervalo $[0.0, 1.0]$.

É importante destacar que as CNNs e RNNs adotadas nas duas arquiteturas descritas são idênticas entre si, com apenas pequenas alterações relativas ao número de elementos de entrada (RNN) ou ao número de camadas até obter a representação adequada da característica de entrada (CNN).

É importante salientar ainda que as diferentes arquiteturas baseadas na fusão antecipada ou tardia são independentes do algoritmo utilizado para identificar as transições em si. Como as redes antecipada e tardia retornam o mesmo conjunto de saída, um algoritmo que seja capaz de realizar a segmentação sobre uma arquitetura pode ser aplicado diretamente na outra arquitetura sem necessitar de qualquer alteração. Essa flexibilidade ocorre também na arquitetura de fusão tardia, no qual o algoritmo de segmentação pode tanto ser utilizado para segmentar a saída obtida da fusão multimodal como também de qualquer rede unimodal individual.

4.4 Características extraídas

Neste trabalho, para instanciar o modelo proposto, foram utilizadas quatro características de três modalidades diferentes consideradas em estado da arte, eficientes ou amplamente adotadas por técnicas reportadas na literatura.

Da modalidade visual, foram extraídas características convolucionais (ConvFeat) e Color-SIFT (CSIFT). As ConvFeat são utilizadas em trabalhos em estado da arte em tarefas diversas relacionadas com processamento de imagens alcançando resultados significativos (RAZAVIAN *et al.*, 2014). Por sua vez, as características CSIFT utilizam o extrator e descritor SIFT com informações de crominância, que adiciona uma alto valor semântico para a identificação de transições de cenas.

É importante destacar que características ConvFeat e SIFT, da qual se baseia o CSIFT, apresentam um elevado grau de complementariedade (YAN *et al.*, 2016; Lv *et al.*, 2018). ConvFeats, desenvolvidas para a classificação, são altamente generalizáveis, sendo capazes de encontrar imagens da mesma classe semântica, embora sejam incapazes de distinguir adequadamente dissimilaridades em imagens de uma mesma classe. Já as características SIFT, criadas para a descrição de pontos relevantes da imagem e caracterizadas pelo seu alto nível discriminativo, são especialmente eficazes em distinguir as diferenças entre imagens inclusive de uma mesma classe. Com o uso de ambas as características, é possível identificar se as tomadas possuem imagens de classes similares (ConvFeat), além de identificar os detalhes e possíveis diferenças entre elas (CSIFT), justificando, assim, o uso de duas características visuais na abordagem proposta.

Para a extração de características visuais ConvFeat e CSIFT é necessário um conjunto de quadros-chave de cada tomada. A seleção de quadros-chaves é geralmente adotada pelos pesquisadores (BARALDI; GRANA; CUCCHIARA, 2015a; LOPES; TROJAHN; GOULARTE,

2014; CHASANIS; LIKAS; GALATSANOS, 2009; SIDIROPOULOS *et al.*, 2011) devido à grande similaridade entre os quadros de uma mesma tomada, sendo que a seleção de alguns quadros específicos resulta em substancial redução no volume de dados a serem processados sem perda substancial de representatividade. Assim, cinco quadros-chave foram selecionados para cada tomada, efetivamente reduzindo o volume de quadros a serem processados e mantendo uma boa relação de quantidade/eficácia na representação da maioria das tomadas da base adotada.

Para realizar a seleção de um conjunto de quadros-chave, dois critérios importantes foram considerados: o tempo de processamento necessário para a seleção em si e a seleção de quadros-chave distintos. Tais aspectos são especialmente importantes pois, caso o processo de seleção seja demasiadamente custoso, reduz-se o eventual ganho de desempenho que a seleção de quadros-chave se objetiva e, caso os quadros-chave selecionados sejam idênticos, podem comprometer a representatividade de tomadas mais complexas. Assim, de maneira a atender tais requisitos, foi adotada a técnica proposta por Trojahn e Goularte (2013).

A técnica de seleção de quadros-chave é baseada na análise de histogramas de cor e no cálculo da dissimilaridade por meio da intersecção entre histogramas dos quadros da tomada. Os quadros da tomada, representados por seus histogramas normalizados HSV 8:4:4, são comparados entre si, de maneira a identificar se tais quadros são "similares". Os quadros-chave são, em ordem, os quadros com o maior nível de similaridade com os demais quadros, desde que não sejam similares com nenhum outro quadro-chave previamente selecionado. A técnica de seleção de quadros-chave empregada, chamada de VKFrameS[1], que seleciona n quadros-chaves de um dado conjunto Q de quadros de entrada de uma tomada, é descrita no Algoritmo 2.

É possível perceber que no Algoritmo 2, Linhas **7** e **19**, *Inter* é a função que retorna o valor de intersecção entre histogramas, com intervalo válido de 0.0 (histogramas diferentes) até 1.0 (histogramas idênticos). Por padrão, a técnica emprega o valor de 0.95 (95% de similaridade) para determinar se dois quadros são considerados "similares", valor este que pode ser alterado para atingir fins específicos, como garantir uma maior dissimilaridade entre os quadros-chave selecionados. O algoritmo procura garantir que os quadros-chave selecionados são representativos da tomada, resultado do cálculo da similaridade entre histogramas (Linhas **7** a **10**) e do uso do índice com maior valor, o mais similar a todos os demais (Linha **15**). Ao comparar um quadro-chave candidato com os quadros-chave já selecionados (Linhas **18** a **22**), o algoritmo visa ainda a assegurar a seleção de quadros-chave únicos ou, especificamente, com ao menos 5% de dissimilaridade quanto a intersecção de seus histogramas.

Com os quadros-chaves selecionados, é realizada a extração de características visuais ConvFeat e CSIFT, aurais MFCC e histogramas *Bag of Words* (BoW) textuais conforme detalhado nas subseções a seguir, respectivamente.

[1] A implementação da VKFrameS pode ser encontrada em <https://github.com/Trojahn/VKFrameS>

Algoritmo 2 – Algoritmo para a seleção de quadros-chave adotado neste trabalho.

Entrada: Os quadros de entrada Q e n, o número desejado de quadros-chave.
Saída: A lista QC, contendo o índice dos quadros-chave selecionados.

$H \leftarrow$ histogramas HSV 8:4:4, normalizados, dos quadros de entrada Q
$i \leftarrow 0$
$Sim \leftarrow$ Lista contendo Q zeros.
enquanto $i <$ tamanho de H **faça**
    $j \leftarrow i+1$
    enquanto $j <$ tamanho de H **faça**
        se $Inter(H[j], H[i]) \geq 0.95$ **então** ▷ $Inter$ é a função de intersecção de histogramas.
            $Sim[j] = Sim[j] + 1$
            $Sim[i] = Sim[i] + 1$
        fim se
    fim enquanto
fim enquanto
$QC \leftarrow$ Lista vazia
enquanto tamanho de $QC < n$ **faça**
    $iMaior \leftarrow$ índice do maior valor de Sim
    $iTemp \leftarrow 0$
    $adicionar \leftarrow$ verdadeiro
    enquanto $iTemp <$ tamanho de QC **faça**
        se $Inter(H[iMaior], H[QC[iTemp]]) \geq 0.95$ **então**
            $adicionar \leftarrow$ falso
        fim se
    fim enquanto
    se $adicionar$ **então**
        Adiciona $iMaior$ em QC
    fim se
    $Sim[iMaior] = 0$
fim enquanto

4.4.1 ConvFeat

Uma característica convolucional (do inglês *convolutional feature* - ConvFeat) é uma característica visual extraída da saída de alguma CNN treinada dado um amplo conjunto de treinamento (SIMO-SERRA *et al.*, 2015). Devido a arquitetura dessas redes, os quadros usados como entrada devem possuir uma dada resolução fixa e predefinida, sendo comum o uso de processos de redimensionamento prévio da imagem.

As ConvFeats são reconhecidas por seu alto poder discriminativo para seu conjunto de treinamento e invariância a diversas transformações visuais como translação e rotação (LECUN *et al.*, 1998; GOODFELLOW; BENGIO; COURVILLE, 2016). Como reportado no trabalho de Razavian *et al.* (2014), as características convolucionais obtém resultados significativos em tarefas e bases de dados diferentes. Tais particularidades tornam as características convolucionais populares em diversas tarefas relacionadas com imagens, como a classificação de imagens e detecção de objetos.

Assim como em trabalhos relacionados na área de segmentação em cenas (BARALDI; GRANA; CUCCHIARA, 2015a; BARALDI; GRANA; CUCCHIARA, 2017), recuperação de conteúdo (VUKOTIC; RAYMOND; GRAVIER, 2018) e classificação de vídeos (SHEN; DEMARTY; DUONG, 2017), foram utilizados vetores de características provenientes de CNNs desenvolvidas para a classificação de imagens, como a VGGNet (SIMONYAN; ZISSERMAN, 2014) ou a ResNet (HE *et al.*, 2016). Tais redes, desenvolvidas para a *ImageNet Large Scale Visual Recognition Competition* (ILSVRC), recebem uma imagem de entrada, classificando-a dentre um conjunto de 1000 classes preestabelecidas.

A extração de características convolucionais, neste trabalho, se baseou em uma rede VGGNet de 16 camadas previamente treinada [2]. Para cada quadro-chave de entrada, redimensionado por meio da biblioteca OpenCV para a resolução requerida de 224x224, definida pelos autores da VGGNet, obtendo assim um um vetor de característica com 4096 dimensões. Assim, como foram utilizados cinco quadros-chave para cada tomada, é obtido um conjunto de 5x4096 vetores de características relativos às características ConvFeat para cada tomada.

4.4.2 CSIFT

A Color SIFT (CSIFT) ou C-color-SIFT é uma característica local que se destaca pois combina informações de cor com as qualidades do SIFT (LOWE, 2004), descrito na Subseção 2.3.1, tal como o alto poder discriminativo e a robustez a diversas transformações geométricas (MIKOLAJCZYK; SCHMID, 2005). Tais particularidades tornam o SIFT extremamente popular em diversas tarefas de vídeo, inclusive na segmentação em cenas (CHASANIS; KALOGERATOS; LIKAS, 2009; LOPES; TROJAHN; GOULARTE, 2014).

A CSIFT, proposta por Burghouts e Geusebroek (2009), se baseia na adição de derivadas gaussianas específicas para cor, desenvolvida por Geusebroek *et al.* (2001), em um descritor SIFT gerando, assim, uma característica geometricamente e fotometricamente invariante. Segundo seus autores, o CSIFT se mostrou mais discriminativo do que o SIFT em imagens com efeitos de sombras e sombreamento, superando-o em bases de dados tais como o Pascal Visual Object Classes Challenge[3] (EVERINGHAM *et al.*, 2015).

Os vetores de características CSIFT foram extraídos por meio da biblioteca OpenIMAJ (HARE; SAMANGOOEI; DUPPLAW, 2011), utilizando o detector de pontos-chave baseado na diferença de gaussianas dos quadros-chave de cada tomada. Cada vetor de característica CSIFT é composto de 364 dimensões, sendo o número de vetores variável conforme as particularidades da imagem como número de cantos T e bordas. Devido a restrições na rede neural e para garantir uma melhor comparação entre tomadas diferentes, o número de vetores de características de cada tomada é padronizado seguindo o seguinte processo (as justificativas do processo adotado

[2] As redes VGGNet treinadas podem ser encontradas em <http://www.robots.ox.ac.uk/vgg/research/very_deep/>

[3] <http://host.robots.ox.ac.uk/pascal/VOC/>

são apresentados na sequência):

1. Para cada tomada, são extraídos os vetores de características CSIFT de seus quadros-chave previamente selecionados.

2. Caso o número de vetores obtido seja superior a 100, é realizado o agrupamento dos vetores encontrados em 100 grupos, sendo escolhido como o representante de cada grupo seu respectivo medóide.

3. Caso o número de vetores seja inferior a 100, os vetores CSIFT da tomada são replicados aleatoriamente até atingir o número de vetores desejado.

O agrupamento utilizado é baseado no algoritmo k-means, usando a distância euclidiana. Após agrupados, são escolhidos os medóides de cada grupo como os vetores de características que representam a tomada. A seleção dos medóides visa a obter um representante daquele conjunto específico de vetores CSIFT, reduzindo o número de vetores ao número previamente definido. O medóide é utilizado para evitar a seleção do centróide que poderia conter valores inválidos para um vetor CSIFT e que poderiam prejudicar a comparação com outras tomadas, que podem ser representadas pelos vetores CSIFT em si (em casos de replicação de vetores, por exemplo). Por sua vez, a replicação de vetores, adotado quando o número de vetores obtido é inferior ao número desejado, não influencia negativamente o resultado na rede neural proposta, devido a adoção de camadas de *max pooling* temporal na CNN desenvolvida, como detalhado na Subseção 4.2.1.

A seleção do medóide de cada grupo é realizado estimando o vetor de característica mais próximo ao centróide do grupo correspondente. Ou seja, dado d um vetor de característica e $centroid_i$ o centróide do i-grupo e $distEucli$ a distância euclidiana entre dois pontos, a seleção do i-medóide é dada por:

$$medoide_i = \min_{d_j \in grupo_i} (distEucli(d_j, centroide_i)) \tag{4.3}$$

Tendo como base de vídeos de avaliação a BBC *Dataset*, são selecionados 100 vetores de características CSIFT por tomada. Tal valor, em testes realizados durante o desenvolvimento, se mostrou adequado, mantendo uma boa relação entre número de vetores de características e eficácia resultante. Portanto, cada tomada é representada por um conjunto de 100x364 vetores de características referentes às características CISFT.

4.4.3 MFCC

Além das características visuais ConvFeat e CSIFT extraídas conforme previamente mencionado, são extraídas características MFCC da modalidade aural. A extração, processo detalhado anteriormente na Subseção 2.3.2, consiste na divisão do áudio do vídeo de acordo com

as transições de tomadas confiáveis. Para cada trecho de áudio, correspondente a uma tomada resultante, um extrator MFCC é executado com áudios a cada 40ms, com janelas deslizantes de 20ms, configuração esta considerada eficaz na BBC *Dataset* segundo método adotado por Baraldi, Grana e Cucchiara (2017). Uma biblioteca auxiliar foi utilizada para a extração dos vetores MFCC[4], usando suas configurações padrão com exceção do tamanho da janela e do deslocamento entre cada janela.

É importante destacar que o áudio utilizado para a extração dos vetores MFCC é de canal único (*monoaural*), com taxa de amostragem de 44100Hz, obtido pela codificação realizada pela ferramenta FFmpeg[5] caso o áudio possua mais que um único canal. O idioma do áudio utilizado para a extração é inglês, opção esta visando a facilitar comparações da técnica com outras reportadas na literatura.

Assim como sua duração, o número de vetores de características MFCC gerado para cada tomada é diferente. Foi adotado, então, um processo idêntico ao descrito na extração das características CSIFT, baseada no agrupamento e seleção de medóides dos grupos, padronizando o número de vetores de características por tomada. Neste trabalho, 500 vetores de características MFCC são extraídos, o que corresponde a 20 segundos de áudio. Na BBC *Dataset* tal valor corresponde a cerca de quatro vezes a duração média das tomadas manualmente segmentadas da base. Por fim, cada tomada da base de vídeos é representada por um conjunto de 500x13 vetores de características referentes as suas características MFCC.

4.4.4 Bag of Words

Por fim, além das características extraídas das modalidades visual (ConvFeat e CSIFT) e aural (MFCC), são extraídas características textuais oriundas das legendas dos vídeos. A importância do uso de tal modalidade é relacionada ao fato de que cada palavra é próxima ao conceito que representa, possuindo elevado nível semântico, o que poderia resultar em melhoras na identificação da transição do assunto ou cena.

Neste trabalho foram utilizadas as legendas no idioma em inglês, com o intuito de facilitar a comparação da abordagem proposta com outras reportadas na literatura. As legendas em si não possuíam qualquer tipo de rótulo, *tag* ou outro metadado qualquer, contendo simplesmente a transcrição da fala do narrador ou personagem. O processo de extração de características BoW adotado, descrito previamente na Subseção 2.3.3, é detalhado a seguir:

1. Com todas as legendas da base de vídeos, são aplicados os processos de *tokenização*, separação das frases da legenda em termos, padronização dos termos para minúsculas e remoção de sinais de pontuação e *stop words*.

[4] A biblioteca para extração de características MFCC pode ser encontrada em <https://github.com/jameslyons/python_speech_features>

[5] A FFmpeg pode ser encontrada em <https://www.ffmpeg.org/>

2. Conversão de cada termo para o seu correspondente vetor de característica Word2Vec usando uma rede previamente treinada.

3. Agrupamento dos vetores de características pelo algoritmo k-means em 200 grupos baseada na distância dos cossenos.

4. Seleção dos medóides de cada grupo, calculado pela distância dos cossenos em relação ao centróide, formando o dicionário textual.

5. Para a legenda da tomada, após os processos 1 e 2, os vetores de características obtidos são comparados com o dicionário textual, gerando um histograma com a ocorrência das palavras. Uma legenda qualquer é considerada pertencente a uma tomada se a mesma ocorre em algum momento da tomada.

6. Cada histograma gerado é normalizado de maneira que a soma de todos os valores em cada índice seja igual a um.

A conversão das palavras extraídas das legendas para um vetor de característica numérico é baseado em uma rede Word2Vec criada por meio de uma biblioteca auxiliar[6] usando seus parâmetros padrões. Para treinamento da rede, uma cópia da Wikipédia na língua inglesa[7] é utilizada, contendo cerca de 5.6 milhões de artigos únicos. Assim como no processo para gerar os histogramas de ocorrências de palavras, todos os sinais de pontuação e as chamadas *stop words* são removidas na fase de pré-processamento dos dados do processo de treinamento da rede Word2Vec.

Após a extração e cálculo dos histogramas de ocorrências de palavras previamente descrito, cada tomada é representada por um histograma normalizado de 200 posições, no qual a soma de todos os índices é igual a um. Nos casos no qual a tomada não possui qualquer legenda, a mesma é representada por um histograma nulo, no qual todos os índices possuem valor zero.

4.5 Segmentação em cenas

Para obter a segmentação em cenas é necessário analisar a saída da rede neural de maneira a identificar as tomadas de transição de cenas. A análise em si, então, constitui um algoritmo de segmentação em cenas. No caso geral, tais algoritmos recebem como entrada representações das tomadas e tem como objetivo encontrar algum padrão ou particularidade que permita a identificação da mudança do assunto e sua consequente transição de cenas. No caso específico deste trabalho, a técnica para identificação de transições de cenas, descrita no Algoritmo 3, é baseada na comparação da classificação tomada-a-tomada obtida pela rede neural com um

[6] A biblioteca *gensim*, adotada para a criação e treinamento da rede Word2Vec, pode ser encontrada em <https://radimrehurek.com/gensim/>

[7] <http://en.wikipedia.org>

limiar de transição/não-transição predefinido. Tal limiar é necessário já que a saída da rede é um valor no intervalo aberto entre 0 e 1, sendo assim necessário definir um valor mínimo para determinar se uma tomada é uma transição de cena. O valor do limiar é diretamente dependente das particularidades de cada vídeo de entrada, no qual a redução de seu valor aumenta o número de transições de cenas detectadas. Em tempos de execução, o valor adequado do limiar pode ser determinado heuristicamente por meio da busca binária, aproximando-se, a cada iteração, do valor adequado para determinado conjunto de vídeos, abordagem esta adotada neste trabalho.

A técnica de comparação tomada-a-tomada foi adotada para avaliar a eficácia de cada rede neural em detectar padrões significativos das características de entrada e correlacioná-las adequadamente, sem depender de técnicas rebuscadas e altamente especializadas de segmentação em cenas. Adicionalmente, a técnica possui diversas particularidades desejáveis: não impõe qualquer tipo de restrição arbitrária para a segmentação em termos de domínio de vídeo, número de cenas ou números de tomada para cada cena; é altamente extensível, podendo ser desenvolvidas diversas melhorias como uma análise em janelas deslizantes para melhor identificar mudanças graduais de contexto/cena; é uma técnica com baixo custo computacional, sendo necessário apenas n comparações simples para segmentar um vídeo com n tomadas.

Algoritmo 3 – Algoritmo de segmentação em cenas baseado na comparação tomada-a-tomada.

Entrada: *vClass*, uma lista de valores gerados pela rede neural para n tomadas de entrada e *limiar*, o valor de limiar usado para identificar tomadas de transição de cena.
Saída: *cenas*, uma lista contendo os índices inicial e final (inclusives) das tomadas que correspondem a uma cena.

```
1: cenas ← lista vazia
2: anterior ← 0
3: para indice faça0n
4:     se vClass[indice] ≥ limiar então
5:         adiciona [anterior, indice] em cenas
6:         anterior ← anterior + 1
7:     fim se
8: fim para
9: cenas ← HeuristicaSubsegmentacao(cenas, vClass)
10: cenas ← HeuristicaSobreSegmentacao(cenas, vClass)
11: cenas ← HeuristicaSegmentacaoPosicao(cenas, vClass)
```

O Algoritmo 3 descrito pode ser aplicado tanto na saída da arquitetura de rede baseada na fusão antecipada ou tardia, como também nas redes unimodais em si, cujas saídas são usadas no processo de fusão tardia. Tal abordagem é baseada na própria modelagem do problema, já que o valor obtido pela rede para cada tomada é a probabilidade de que tal tomada seja realmente de transição. Assim, ao definir um valor de probabilidade mínima (a variável de entrada *limiar* no Algoritmo 3), procede-se com a binarização da saída da rede e, por consequência, a obtenção da segmentação em cenas.

Normalmente, técnicas de segmentação em cenas (RASHEED; SHAH, 2003; CHA-

SANIS; KALOGERATOS; LIKAS, 2009; TROJAHN; GOULARTE, 2013) podem produzir comportamentos que reduzem sua eficácia. No caso deste trabalho alguns comportamentos possíveis são:

- Algumas tomadas de transição podem não atingir o limiar mínimo predeterminado para a identificação da transição de cenas, embora seu valor seja significativamente superior aos valores obtidos em tomadas vizinhas, o que pode indicar que tais tomadas sejam realmente de transição. A falha da detecção de transição, nestes casos, é fruto do baixo valor da classificação da rede neural aliado ao algoritmo de identificação de cenas (Algoritmo 3) que não considera as tomadas adjacentes na análise.
- Uma tomada de transição que apresenta alto valor de saída pode ser seguida de outra(s) tomada(s) também de valor consideravelmente elevado, embora não sejam tomadas de transição. Ou seja, a presença de uma tomada de transição de alto valor pode tornar a próxima tomada também de transição, resultando em altos índices de sobre-segmentação.
- Algumas tomadas podem ser identificadas como de transição apenas devido a sua posição relativa, tal como a primeira tomada ou a última tomada do vídeo.

Assim, para prevenir o impacto de tal comportamento na segmentação, três heurísticas simples foram desenvolvidas, sendo duas especificamente para reduzir a sub-segmentação e uma para reduzir a sobre-segmentação. A descrição de cada heurística aplicada é apresentada a seguir:

- Nos casos nos quais duas tomadas identificadas como de transição de cenas estejam distantes por um número predefinido de tomadas (sub-segmentação), o maior valor de saída da rede obtido é selecionado para se tornar uma nova borda de cena.
- Em uma janela deslizante de três tomadas de tamanho, caso a tomada central apresente um valor de saída significativamente acima das vizinhas, mesmo que abaixo do limiar usado para identificar as tomadas de transição, tal tomada será considerada como de transição.
- Caso um número predeterminado de tomadas consecutivas sejam classificadas como de transição, a tomada que apresentar o menor valor de saída será, então, classificada como uma tomada de não-transição, efetivamente reduzindo a sobre-segmentação em tais trechos.

As heurísticas para remoção de casos de sub-segmentação por valores de classificação abaixo do limiar predefinido, dos casos de sobre-segmentação em diversas tomadas adjacentes classificadas como de transição e de casos no qual a transição ocorre apenas devido a posição temporal no vídeo, representadas pelas funções *HeuristicaSubsegmentacao*, *HeuristicaSobreSegmentacao* e *HeuristicaSegmentacaoPosicao*, respectivamente, são representadas nas linhas **8** a **10** no Algoritmo 3.

4.6 Discussões sobre o capítulo

Neste capítulo foram descritos o método de segmentação proposto, formado da modelagem do problema de segmentação em cenas como um problema de classificação de tomadas, além de um modelo. O modelo, por sua vez, é formado por redes convolucionais, responsáveis por obter representações adequadas das tomadas de entrada, além de redes recorrentes, capazes de analisar o correlacionamento temporal para classificar as tomadas. A seguir, foram descritas duas arquiteturas de redes neurais seguindo o modelo proposto, utilizando a abordagem de fusão tardia ou antecipada. A extração de características utilizadas para instanciar as arquiteturas desenvolvidas foi descrita, formada de quatro características consideradas em estado da arte em três modalidades diferentes. Por fim, o algoritmo de segmentação em cenas, responsável por identificar as tomadas de transição de cenas, foi descrito.

A modelagem da tarefa de segmentação em cenas como um problema de classificação de tomadas proposta permite que uma ampla gama de vídeos possa ser segmentada, independentemente do seu domínio e de suas particularidades como tamanho das cenas ou duração total. Diferentemente da abordagem baseada em agrupamento, adotada por Baraldi, Grana e Cucchiara (2015a) em um trabalho relacionado que também utiliza Aprendizagem Profunda, por exemplo, a modelagem não requer o custoso processo de agrupamento de tomadas.

O modelo proposto, composto de CNNs e RNNs, é caracterizado pela sua alta flexibilidade quanto as características de entrada e a abordagem de fusão multimodal a ser utilizada. Por meio da alteração do número de camadas e neurônios na rede convolucional, uma ampla gama de características de entrada pode ser analisada de maneira a gerar uma representação, formada de um vetor de característica único, que pode ser analisada e classificada por uma rede recorrente. Já a rede recorrente, devido sua flexibilidade quanto a natureza de sua entrada, é capaz de identificar padrões temporais em diferentes configurações, suportando tanto representações unimodais como multimodais. Tais particularidades tornam o modelo apto a ser adotado com abordagens de fusão antecipada ou tardia, gerando assim duas arquiteturas de rede similares.

Para realizar a segmentação em cenas, após a execução da rede neural pré-treinada, é necessário realizar a identificação das tomadas de transição. Nesse sentido, o algoritmo de segmentação utilizado é consequência direta da definição utilizada na modelagem do problema proposto. Comparando o valor obtido tomada-a-tomada com um limiar predefinido, a técnica possui baixo custo computacional aliado à sua independência quanto ao domínio ou particularidades do vídeo de entrada. É importante destacar que a escolha de tal algoritmo para segmentação em cenas é devido ao foco deste trabalho na proposta do modelo e não na técnica de segmentação em si. O uso de uma abordagem mais rebuscada e refinada, resultando em melhores segmentações, poderia ofuscar a eficácia do modelo proposto.

O método de segmentação proposto, utilizando a modelagem do problema proposta e duas arquiteturas de rede desenvolvidas que formam o modelo, foi avaliado na tarefa da

segmentação temporal de vídeo em cenas, conforme descrito no Capítulo 5.

CAPÍTULO

5

AVALIAÇÃO

No capítulo anterior, descreveu-se o método proposto utilizando uma modelagem do problema de segmentação em cenas como de classificação de tomadas usando, para isso, duas arquiteturas de redes convolucionais e recorrentes baseadas na fusão antecipada ou tardia. O método proposto, como reportado, provê uma alta flexibilidade da extração e representação de características com o processo de segmentação em si. Além disso, o método suporta a escolha entre as abordagens de fusão multimodal antecipada ou tardia de maneira transparente e, novamente, flexível quanto as características de entrada.

Assim, este capítulo tem o intuito de avaliar a eficácia do método proposto, que supõe prover uma melhor análise do correlacionamento temporal entre tomadas adjacentes devido ao uso de redes convolucionais e recorrentes. A avaliação da segmentação em cenas foi realizada com as métricas adotadas por pesquisadores da área em uma base de dados publicamente disponível. Considerações sobre o desempenho também são apresentadas, medindo o tempo necessário para o treinamento e a obtenção da segmentação em cenas das redes neurais desenvolvidas.

O restante deste capítulo é assim dividido: A Seção 5.1 discute o processo de treinamento das redes neurais utilizadas, no qual os pesos das conexões ou sinapses entre os elementos ou neurônios da rede são ajustados de maneira a aprender a identificar as tomadas de transições de cenas; Já a metodologia dos testes realizados ao comparar a segmentação obtida pelas redes neurais em uma base de vídeos pública é detalhada na Seção 5.2; Os resultados obtidos na avaliação usando as diferentes configurações de fusão multimodal, antecipada e tardia, são apresentados na Seção 5.3; Por sua vez, a Seção 5.4 apresenta a comparação dos resultados de eficácia obtidos com técnicas em estado da arte presentes na literatura, evidenciando a eficácia do método proposto frente ao estado da arte atual, além de considerações sobre desempenho entre as diferentes abordagens; Finalmente, a Seção 5.5 apresenta uma discussão sobre este capítulo.

5.1 Treinamento das redes desenvolvidas

Para avaliar a eficácia das redes desenvolvidas, é necessário realizar o treinamento das redes neurais propostas com o intuito de ajustar os pesos ou parâmetros da rede, obtendo uma classificação adequada da cena e aprendendo os padrões que determinarão as transições de cenas.

Nesse sentido, o treinamento foi realizado baseado na técnica *leave-one-out* (EFRON, 1987), também adotado em trabalhos relacionados (PETER *et al.*, 2014; BARALDI; GRANA; CUCCHIARA, 2015a; BARALDI; GRANA; CUCCHIARA, 2017). Nessa abordagem, todos os vídeos da base de dados, com exceção daquele que será usado para avaliar a eficácia da rede, são usados para o treinamento. Tal abordagem foi utilizada pois 1) providencia um maior número de dados de treinamento, especialmente importante em bases de dados com número reduzido de amostras de entrada como a BBC *Dataset* e 2) o uso de outras configurações de validação cruzada (do inglês *cross validation*) poderia acarretar em um viés dependendo da escolha do conjunto de vídeos de treinamento/validação.

Para o processo de treinamento em si, utilizou o algoritmo do gradiente descendente estocástico (do inglês *stochastic gradient descendent* - SGD) foi utilizado. Tal escolha é justificada pela sua ampla popularidade e por ser considerado um algoritmo eficiente e eficaz de otimização em diversas tarefas de aprendizagem de máquina (KINGMA; BA, 2014), inclusive na segmentação em cenas (BARALDI; GRANA; CUCCHIARA, 2015a; BARALDI; GRANA; CUCCHIARA, 2017).

Neste trabalho, como o objetivo das redes desenvolvidas é classificar uma dada tomada de entrada em um valor binário, foi utilizada a função de perda (do inglês *loss function*) de entropia binária cruzada (do inglês *binary cross entropy* - BCE) descrita na Equação 5.1. É importante destacar que a escolha da BCE é devido a restrição imposta pela própria função linear sigmoidal utilizada após a rede recorrente (Equação 4.2), cuja saída é um valor único no intervalo aberto entre 0 e 1.

$$\mathscr{L}(o,t) = -\frac{1}{n} \cdot \sum_{i}^{n} t_i * \log(o_i) + (1-t_i) \cdot \log(1-o_i) \tag{5.1}$$

Onde o é o vetor de saída obtido da rede neural e t é o vetor de resultados esperados (obtido da segmentação em cenas confiável), ambos de tamanho n. Note que t é composto de valores binários (0 ou 1), sendo cada valor o_i interpretado como a probabilidade de $t_i = 1$, ou seja, $o_i \in]0,1[$.

Dois critérios diferentes de parada do algoritmo SGD foram adotados: quando o erro médio da função de perda do treinamento ficar abaixo de um limiar predefinido, ou quando o número máximo predefinido de iterações (épocas) for alcançado. Ambos os critérios são utilizados para determinar se a rede foi capaz de convergir e que, em caso positivo, deve encerrar o treinamento. Nas redes treinadas neste trabalho, foi utilizado como limiar de erro médio o valor

de 0.01, que indica que a rede é capaz de classificar os dados de treinamento com alta precisão, além de um número máximo de iterações de 600 épocas, um valor considerado adequado haja visto as particularidades da tarefa e das características de entrada utilizadas.

É importante mencionar que algumas redes podem não convergir ao final do número máximo de épocas preestabelecido. Tal fenômeno ocorre devido, principalmente, a aleatoriedade dos pesos ao instanciar a rede a ser treinada, associada a tendência do algoritmo SGD de atingir um mínimo local. Esse problema ocorre principalmente em redes com maior número de valores de entrada (maior dimensionalidade), ou que possuem maior número de unidades, tal como as redes seguindo a abordagem de fusão antecipada. Nos casos específicos em que a rede não convergiu, o treinamento foi repetido, utilizando um conjunto de pesos iniciais diferente, até que o algoritmo convergisse. Como critério para determinar se a rede convergiu ou não, foi adotado o limiar de 0.1 de erro médio da função de perda, indicando um baixo índice de erro da rede ao classificar os dados de entrada. Ou seja, se ao final da 600ª época o erro médio estiver acima de 0.1, o treinamento é reiniciado. Tais limiares foram empiricamente determinados de maneira a permitir um treinamento eficiente e com bons resultados de eficácia.

A técnica do anelamento (do inglês *annealing*) (ROBBINS; MONRO, 1951) da taxa de aprendizagem em cada época foi utilizado, tanto para obter uma maior velocidade de treinamento, como reduzir a possibilidade do algoritmo atingir um mínimo local. Assim, conforme o número de épocas que o algoritmo já executou, uma nova taxa de aprendizagem, de menor magnitude, é adotada. Neste trabalho, as taxas de aprendizagem usadas foram $[2, 0.5, 0.1, 0.05, 0.01]$, nos quais os dois primeiros valores são especialmente elevados para reduzir o impacto da aleatoriedade dos valores iniciais dos pesos da rede e acelerar a convergência da rede, seguido de valores consideravelmente menores para garantir um ajuste mais fino ou detalhado dos pesos. A taxa de aprendizagem é alterada (do maior para o menor) a cada 75 épocas, com exceção da última taxa de aprendizagem especificada que é utilizada até o fim do treinamento. Tal configuração se mostrou adequada ao treinamento de todas as redes propostas para a segmentação em cenas do BBC *Dataset*, descrita na Subseção 2.6.1, tanto no caso da fusão antecipada quanto nas redes presentes na fusão tardia.

É importante ressalvar o desbalanceamento dos dados de treinamento percebido, resultado inerente do problema da segmentação em cena. Dado que o número de tomadas de um vídeo é geralmente muito superior ao número de cenas, há um expressivo número de tomadas que não são de transição frente a poucas tomadas que são de transição. Uma consequência imediata de tal fato é que as redes treinadas, em suas primeiras épocas, tendem a classificar toda e qualquer tomada como não sendo de transição, independentemente das características de entrada. Outra consequência é a de que, mesmo após a convergência do algoritmo, os valores obtidos para cada tomada se aproximam do limiar inferior do intervalo $[0.0, 1.0]$, mesmo àquelas que podem ser classificadas como tomadas de transição. Tais resultados, porém, não invalidam o treinamento, haja visto que os resultados da eficácia resultante não foram significativamente impactados.

5.2 Metodologia

Os testes foram realizados sobre a BBC *Dataset* usando as redes neurais treinadas seguindo a metodologia de treinamento previamente mencionado, medindo a eficácia da segmentação em cenas resultantes por meio das métricas Precisão/Abrangência e Cobertura/Transbordamento.

A Precisão (P), Abrangência (R) e sua medida F-measure (F_{PR}) foram calculadas com um limiar de tolerância de até três tomadas (HANJALIC; LAGENDIJK; BIEMOND, 1999). Assim, uma transição detectada será considerada verdadeira positiva se estiver em até três tomadas de distância da correspondente transição confiável. Tal abordagem pode ser considerada válida devido a sua utilização em trabalhos relacionados, o que facilita a sua comparabilidade. Por sua vez, a Cobertura (C), Transbordamento (O) e sua medida F-measure (F_{CO}) foram calculados por meio da implementação de tais métricas, sem qualquer alteração, disponibilizado por Baraldi, Grana e Cucchiara (2015a), aumentando a confiabilidade da comparação entre técnicas. A métrica NO foi implementada de acordo com a Equação 2.24.

O limiar utilizado para determinar se uma dada tomada é de transição, requisito da técnica de segmentação conforme descrito no Algoritmo 3, foi definido empiricamente. Para cada possível valor de limiar no intervalo $[0.05, 0.5]$, com incrementos de 0.05, foi selecionado o limiar que resulte no melhor valor médio nas métricas F_{PR}, F_{CO} ou F_{CNO}, respeitando o requisito de que o F_{PR} para cada vídeo seja acima de 50% para evitar que casos extremos de sub-segmentação afetem a métrica de Transbordamento e, consequentemente, a métrica F_{CO}. Além da escolha do limiar de transição de cenas, para evitar que casos de sobre-especialização no treinamento afetem negativamente os resultados, a segmentação obtida de cada rede individual (fusão tardia ou antecipada) foi extraída a cada cinco épocas. Foi selecionada, de maneira exaustiva, a época da rede que obteve o melhor resultado, com o respectivo limiar, na correspondente métrica analisada. No caso particular da fusão tardia, foram selecionadas as redes unimodais cujas épocas obtém o melhor valor de F_{CNO} para o treinamento de sua rede de fusão, considerando o limiar 0.2, valor este que apresentou os melhores resultados na métrica citada. Como reportado nos resultados obtidos, a F_{CNO} apresentou um equilíbrio entre as métricas F_{CO} e F_{PR}, justificando, assim, sua escolha para tal propósito.

Os resultados são apresentados ao priorizar os resultados médios da F_{PR}, F_{CO} ou F_{CNO}. Tal metodologia permite uma comparação mais justa com trabalhos relacionados caso os autores optem por uma métrica específica. Além disso, como medem diferentes aspectos da segmentação, a apresentação dos resultados de tal forma permite expor o melhor caso de cada rede em particular e a identificação dos limites particulares da abordagem ou configuração adotada.

É importante destacar que, ao priorizar determinada métrica, uma segmentação resultante potencialmente diferente das obtidas em outras métricas é avaliada, haja visto que o resultado ótimo para tal métrica pode ocorrer em uma época de treinamento ou limiar de identificação de

transições de cenas distintos. Por exemplo, as segmentações selecionadas que apresentam melhor F_{PR} médio podem ter alto valor de sobre-segmentação, enquanto que as segmentações selecionadas ao priorizar a F_{CO} podem apresentar alto valor de sub-segmentação, mesmo utilizando um mesmo limiar de identificação de cenas.

Os valores das métricas obtidos para cada vídeo, inclusive suas médias, foram arredondados para o número inteiro mais próximo. O desvio padrão, por sua vez, foi truncado na segunda casa decimal.

5.3 Resultados obtidos

Os resultados obtidos pelas redes baseadas na fusão antecipada e tardia, assim como uma discussão acerca do desempenho e eficácia entre as duas arquiteturas, são apresentados nas Subseções a seguir.

5.3.1 Fusão antecipada

O melhor resultado médio de F_{PR} na arquitetura baseada na fusão antecipada foi obtido com o limiar de identificação de cenas no valor de 0.15. Os resultados individuais para tal configuração são descritos na Tabela 6.

Tabela 6 – Resultados obtidos pela rede neural desenvolvida baseada na fusão antecipada ao priorizar a métrica de F_{PR}. O limiar de identificação de cenas utilizado é igual a 0.15.

	P	R	F_{PR}	C	O	F_{CO}	NO	F_{CNO}
From Pole to Pole	68	76	71	72	62	50	46	62
Mountains	58	74	65	72	41	65	39	66
Ice Worlds	60	79	68	70	49	59	39	66
Great Plains	65	77	71	81	49	63	41	68
Jungles	56	82	66	68	40	64	36	66
Seasonal Forests	56	89	69	70	31	69	28	71
Fresh Water	60	60	60	76	61	51	50	60
Ocean Deep	66	72	68	76	37	69	42	66
Shallow Seas	68	85	76	75	43	65	37	69
Caves	63	70	67	68	50	57	42	63
Deserts	60	71	65	72	34	69	36	68
Média	62	76	68	73	45	62	40	66
Desvio Padrão	4.52	8.08	4.18	4	10.14	6.75	5.84	3.2

Ao priorizar o F_{PR} médio, que mede o quanto as tomadas de transição foram corretamente detectadas, é perceptível um elevado grau de Abrangência (R), especialmente em vídeos como **Seasonal Forests** (89%) e **Shallow Seas** (85%), indicando uma alta eficácia da referida arquitetura de detectar a maioria das transições de cenas presentes na base de vídeos. Concomitantemente, o alto valor de Cobertura (C) indica que partes significativas das cenas confiáveis

foram corretamente agrupadas em suas respectivas cenas detectadas. Embora o valor de Precisão seja abaixo da Abrangência, não foram detectados índices expressivos de sobre-segmentação, fato também apoiado pelos valores significativos de Transbordamento (O) e de Cobertura (C).

Os resultados obtidos nessa configuração indicam uma significativa dependência das informações visuais presentes nos vídeos. O vídeo **Fresh Water** possui como particularidade ter tomadas adjacentes com alto índice de similaridade visual, nas quais a flora e fauna representada possuem diversas características em comum, que dificulta a identificação das trocas de cenas entre os vídeos. Como resultado, tal vídeo obteve os piores resultados em todas as métricas utilizadas, com especial destaque para o Transbordamento (O) muito acima da média dos outros vídeos analisados. Nesse caso específico, a alta similaridade visual entre as tomadas fez com que a rede neural fosse incapaz de detectar a maioria das transições de cena do vídeo. Seu valor de Precisão (P), próximo da média, indica que uma significativa parcela das transições que apresentavam alta dissimilaridade visual foram corretamente detectadas.

O vídeo **From Pole to Pole** exibe alguns animais dos polos Sul e Norte e de alguns fatos sobre seus estilos de vida, apresentando um relativo grau de similaridade visual devido ao plano de fundo, mas com claras distinções aurais entre os diferentes animais e transições de cenas bem demarcadas, resultando em valores de Precisão (P) e Abrangência (R) acima da média. A métrica de Transbordamento (O), por outro lado, apresenta um valor anormalmente elevado, considerado como *outlier* já que as demais métricas estão próximas da média. Nesse caso em particular, determinadas transições de cenas com alta similaridade visual e aural (tomadas sem narração) não foram detectadas pela rede neural, ocasionando certa sub-segmentação localizada, que levou a métrica a retornar valores inválidos e incorretamente elevados. A formulação alternativa da métrica de Transbordamento (NO), criada para melhor representar casos de sub-segmentação localizadas e evitar tal limitação, indica um valor correto e mais próximo da média.

Por sua vez, o melhor resultado médio de F_{CO} na arquitetura baseada na fusão antecipada foi obtido com o limiar de identificação de cenas no valor de 0.2. Os resultados individuais para tal configuração são descritos na Tabela 7.

Os resultados ao priorizar a métrica de F_{CO}, que busca medir o quanto as tomadas das cenas foram corretamente agrupadas, indicam um cenário diferente do obtido anteriormente. Primeiro, há o surgimento de casos expressivos tanto de sub-segmentação (**Deserts**, **Caves** e **Mountains**) como de sobre-segmentação (**From Pole to Pole**, **Ice Worlds**, **Great Plains** e, especialmente, **Fresh Water**) baseados nos resultados de Precisão (P) e Abrangência (R). Segundo, há uma queda expressiva no valor médio de Transbordamento (O) que, mantido o valor de Cobertura (C), eleva os resultados nas métricas de F_{CO} e F_{CNO}.

Nesta avaliação, a similaridade do valor obtido no Transbordamento (O) em cada vídeo revela uma certa homogeneidade nos vídeos semanticamente mais relacionados. Nesse sentido os vídeos **Jungles**, **Seasonal Forests** e **Great Plains**, além de valores idênticos de Transbordamento (O), consistem em apresentar ambientes semanticamente relacionados com relativa similaridade

Tabela 7 – Resultados obtidos pela rede neural desenvolvida baseada na fusão antecipada ao priorizar a métrica de F_{CO}. O limiar de identificação de cenas utilizado é igual a 0.2.

	P	R	F_{PR}	C	O	F_{CO}	NO	F_{CNO}
From Pole to Pole	44	83	58	60	37	61	29	65
Mountains	65	42	51	82	36	72	41	69
Ice Worlds	50	84	63	68	32	68	28	70
Great Plains	59	86	70	76	38	68	34	71
Jungles	55	75	64	66	38	64	34	66
Seasonal Forests	49	77	60	72	38	67	30	71
Fresh Water	37	90	53	55	29	62	28	62
Ocean Deep	62	62	62	77	23	77	41	67
Shallow Seas	72	79	75	77	37	69	31	73
Caves	63	42	51	83	34	74	70	44
Deserts	55	47	51	78	23	78	51	60
Média	56	70	60	72	33	69	38	65
Desvio Padrão	9.96	18.31	8.2	8.98	5.87	5.54	12.76	7.94

visual em relação a flora, como florestas e gramíneas em geral, consequentemente obtendo segmentações com eficácia similar em tais vídeos. Esse comportamento também aparece em outro conjunto de vídeos semanticamente relacionados, formado por **From Pole to Pole**, **Mountains** e **Ice Worlds**, que mostram a fauna em ambientes relativamente similares, formado por regiões geladas com predominância de gelo e neve. O vídeo **Ice Worlds**, por sua vez, possui um **Transbordamento** (O) relativamente dissimilar aos outros dois vídeos pois contém um trecho significativamente longo em ambiente noturno e escuro, particularidade praticamente ausente nos demais vídeos.

Novamente, a relevância visual para a identificação da segmentação em cenas é exacerbada nesta configuração. Os vídeos **Ocean Deep** e **Deserts** possuem baixa relação semântica, no qual o primeiro exibe informações de fauna de fossas oceânicas e o segundo relaciona-se com desertos em ambientes extremamente áridos. Embora semanticamente dissimilares, ambos possuem como particularidade visual uma certa ofuscamento e dificuldade de identificação dos animais sendo apresentados, seja pela presença de ambientes escuros e altamente similares entre si (**Ocean Deep**), seja pela presença de tempestades de areia e alta similaridade entre a cor de pele dos animais e do ambiente de fundo (**Deserts**). Tais vídeos, embora semanticamente dissimilares, apresentam resultados de Cobertura (C) e Transbordamento (O) similares, o que pode indicar a capacidade da rede neural de tratar igualmente vídeos com diferentes contextos semânticos, mas com certas particularidades de gravação e edição em comum.

Por sua vez, o melhor resultado médio de F_{CNO} na arquitetura baseada na fusão antecipada foi obtido com o limiar de identificação de cenas no valor de 0.15. Os resultados individuais para tal configuração são descritos na Tabela 7.

Ao priorizar a métrica F_{CNO}, as segmentações obtidas pela rede neural apresentam um

Tabela 8 – Resultados obtidos pela rede neural desenvolvida baseada na fusão antecipada ao priorizar a métrica de F_{CNO}. O limiar de identificação de cenas utilizado é igual a 0.15.

	P	R	F_{PR}	C	O	F_{CO}	NO	F_{CNO}
From Pole to Pole	63	76	69	72	53	57	39	66
Mountains	42	92	58	59	18	69	19	69
Ice Worlds	50	85	63	68	31	69	27	71
Great Plains	47	92	62	66	27	70	22	72
Jungles	50	86	64	62	34	64	30	66
Seasonal Forests	56	89	69	70	31	69	28	71
Fresh Water	43	81	56	61	37	62	33	64
Ocean Deep	53	77	63	74	41	66	34	69
Shallow Seas	57	89	70	71	32	70	27	72
Caves	48	82	61	69	33	68	30	69
Deserts	47	78	59	67	27	70	25	71
Média	51	84	63	67	33	67	29	69
Desvio Padrão	6.35	5.84	4.57	4.63	8.84	4.15	5.74	2.66

equilíbrio entre os valores obtidos nos dois testes anteriores. Ao contrário do percebido quando a F_{CO} foi priorizada, as segmentações resultantes não apresentam sub-segmentação, ocasionando um aumento expressivo no valor médio de Abrangência (R) e, por consequência, do valor médio de F_{PR}. Os resultados indicam ainda o aumento no número de vídeos com sobre-segmentação, embora não estejam presentes casos extremos como o vídeo **Fresh Water** ao priorizar a F_{CO} (Tabela 7).

Assim como na métrica F_{CO}, os resultados priorizando o valor médio de F_{CNO} buscam encontrar segmentações que apresentem equilíbrio entre o número de tomadas agrupadas em uma mesma cena sem exceder os limites das cenas adjacentes. Nesse caso, as segmentações obtidas priorizam encontrar um maior número de transições de maneira a reduzir significativamente o número de tomadas agrupadas em cenas incorretas (NO). Por consequência, uma maior sensibilidade da rede neural foi percebida, que passa a identificar tomadas de transição mesmo entre tomadas relativamente similares. Em outras palavras, a segmentação ao priorizar a F_{CNO} agrupa apenas as tomadas com alta similaridade, identificando as demais tomadas como transições de cenas.

5.3.2 Fusão tardia

O melhor resultado médio de F_{PR} na arquitetura baseada na fusão tardia foi obtido com o limiar de identificação de cenas no valor de 0.05. Os resultados individuais para tal configuração são descritos na Tabela 9.

Os resultados reportados na Tabela 9 indicam a presença de casos de sobre-segmentação, especialmente em vídeos como **Fresh Water**, **Caves** e **Mountains**. As métricas de Cobertura (C) e Transbordamento (O), que apresentaram quedas acentuadas no valor médio quando com-

Tabela 9 – Resultados obtidos pela rede neural desenvolvida baseada na fusão tardia ao priorizar a métrica de F_{PR}. O limiar de identificação de cenas utilizado é igual a 0.05.

	P	R	F_{PR}	C	O	F_{CO}	NO	F_{CNO}
From Pole to Pole	54	86	66	61	36	62	31	64
Mountains	54	94	68	58	33	62	29	64
Ice Worlds	57	79	66	67	39	64	35	66
Great Plains	57	82	67	67	36	65	33	67
Jungles	51	82	63	63	42	60	32	65
Seasonal Forests	59	69	64	69	32	69	41	64
Fresh Water	46	81	59	61	37	62	32	64
Ocean Deep	52	89	65	61	28	66	25	67
Shallow Seas	66	79	72	64	47	58	38	63
Caves	48	89	62	60	32	64	28	65
Deserts	57	80	67	69	38	65	35	67
Média	55	83	65	64	36	63	33	65
Desvio Padrão	5.55	6.73	3.45	4.08	5.33	2.93	4.62	1.48

parados a arquitetura de fusão antecipada, indicam que uma menor porção das tomadas foram corretamente agrupadas, reforçando os indícios de casos de sobre-segmentação.

Assim como os resultados obtidos na arquitetura de fusão antecipada (Tabela 6), o vídeo **Fresh Water** obteve um dos piores resultados em todas as métricas utilizadas, o que pode indicar que as particularidades intrínsecas do vídeo e não a abordagem de fusão multimodal possuem maior relevância na eficácia resultante.

Há uma discrepância entre as métricas F_{PR} e F_{CO}, perceptível no vídeo **Shallow Seas**. Tal vídeo obteve, ao mesmo tempo, o melhor índice de F_{PR} (72%) com um equilíbrio entre sub e sobre-segmentação e o pior F_{CO} (58%) entre os valores avaliados. A referida discrepância também ocorre entre a F_{PR} e F_{CNO}, no qual também obteve seu pior resultado (63%), embora consideravelmente em menor magnitude. Tal fato indica que a correta identificação dos momentos de transição (Precisão/Abrangência) pode não coincidir com a avaliação da quantidade de tomadas corretamente agrupadas em uma mesma cena (Cobertura/Transbordamento), reforçando a importância de utilizar ambas as métricas para melhor avaliar o comportamento do método de segmentação em cenas proposto.

Por sua vez, o melhor resultado médio de F_{CO} na arquitetura baseada na fusão tardia foi obtido com o limiar de identificação de cenas no valor de 0.05. Os resultados individuais para tal configuração são descritos na Tabela 10, indicando um aumento no número de casos de sobre-segmentação, como percebido em vídeos como **Mountains** e **Caves**, como também percebido na fusão antecipada ao priorizar o valor médio de F_{CO}.

Diferentemente dos resultados da fusão antecipada (Tabela 7, porém, não há mudanças significativas em todas as métricas utilizadas. Um leve decréscimo nos valores médios de Precisão e Abrangência foi percebido, o que resulta em uma queda de cerca de 1% apenas no valor médio

Tabela 10 – Resultados obtidos pela rede neural desenvolvida baseada na fusão tardia ao priorizar a métrica de F_{CO}. O limiar de identificação de cenas utilizado é igual a 0.05.

	P	R	F_{PR}	C	O	F_{CO}	NO	F_{CNO}
From Pole to Pole	52	88	66	60	33	63	29	65
Mountains	51	98	67	58	30	63	26	65
Ice Worlds	57	79	66	67	39	64	35	66
Great Plains	57	80	67	68	36	66	33	67
Jungles	51	82	63	63	42	60	32	65
Seasonal Forests	56	52	54	75	28	73	52	59
Fresh Water	46	81	59	61	37	62	32	64
Ocean Deep	54	74	62	64	24	69	32	66
Shallow Seas	59	82	68	61	42	59	35	63
Caves	48	89	62	60	32	64	28	65
Deserts	55	82	66	68	37	66	34	67
Média	53	81	64	64	35	64	33	65
Desvio Padrão	4.03	11.37	4.26	5.12	5.73	4.02	6.71	2.29

de F_{PR}. Já os valores de Cobertura (C), Transbordamento (O e NO) permanecem com variações abaixo de 1%. Tais resultados indicam que a mudança do limiar de transições de cenas ou a priorização de uma outra métrica têm baixa influência no resultado na arquitetura de fusão antecipada.

Por fim, o melhor resultado médio de F_{CNO} na arquitetura baseada na fusão tardia foi obtido com o limiar de identificação de cenas no valor de 0.05. Os resultados individuais para tal configuração são descritos na Tabela 11.

Tabela 11 – Resultados obtidos pela rede neural desenvolvida baseada na fusão tardia ao priorizar a métrica de F_{CNO}. O limiar de identificação de cenas utilizado é igual a 0.05.

	P	R	F_{PR}	C	O	F_{CO}	NO	F_{CNO}
From Pole to Pole	52	88	66	60	33	63	29	65
Mountains	51	98	67	58	30	63	26	65
Ice Worlds	57	79	66	67	39	64	35	66
Great Plains	57	80	67	68	36	66	33	67
Jungles	51	82	63	63	42	60	32	65
Seasonal Forests	56	72	63	68	28	70	40	64
Fresh Water	46	81	59	61	37	62	32	64
Ocean Deep	52	89	65	61	28	66	25	67
Shallow Seas	65	79	71	64	46	59	38	63
Caves	48	89	62	60	32	64	28	65
Deserts	57	80	67	69	38	65	35	67
Média	54	83	65	63	35	64	32	65
Desvio Padrão	5.27	7.04	3.23	4	5.87	2.97	4.79	1.35

Já os resultados obtidos pela rede baseada na fusão tardia ao priorizar a métrica de F_{CNO} são descritos na Tabela 11. É notável que a métrica teve variação abaixo de 1% nos três testes

realizados (Tabelas 9, 10 e 11), sendo perceptível apenas a variação no seu desvio padrão (melhor caso ao priorizar a F_{CNO} e pior caso ao priorizar a F_{CO}).

5.3.3 Comparação entre as arquiteturas

A avaliação de eficácia realizada sobre as duas arquiteturas, uma baseada na fusão antecipada e a outra na fusão tardia, indicam uma leve vantagem da primeira perante a última, devido a diferenças significativas obtidas nas métricas de F_{CO}. Ao priorizar tal métrica, especialmente utilizada em trabalhos recentes reportados na literatura, a arquitetura baseada na fusão antecipada superou a arquitetura baseada na fusão tardia em 8 dos 11 vídeos da base de vídeos utilizada, exceto nos vídeos **From Pole to Pole** e **Seasonal Forests**, com empate no vídeo **Fresh Water** que obteve o pior resultado individual em ambas as arquiteturas. Tal resultado vai ao encontro das conclusões reportadas na literatura de que, quando um bom modelo pode ser obtido, a fusão antecipada apresenta resultados superiores aos obtidos com a fusão tardia. Neste trabalho a fusão antecipada mostrou uma vantagem, na média, de cerca de 5% de F_{CO} sobre a fusão tardia.

Quanto ao processo de treinamento, quando comparada a abordagem baseada na fusão antecipada, a rede baseada na fusão tardia apresenta as seguintes características:

- Cada rede unimodal requer um menor número de épocas para convergir. Tal resultado é consequência da ausência de modalidades extras no treinamento que podem criar padrões divergentes que demandam maior esforço para serem aprendidos.

- O tempo de processamento de cada época é inferior ao tempo de processamento necessário na rede baseada na fusão antecipada. Este comportamento já era esperado, haja visto o menor volume de dados e de neurônios e pesos na rede neural que cada rede unimodal possui.

Para obter tais conclusões, foi realizado uma avaliação da eficiência ou desempenho das diferentes abordanges propostas. Para isso, foi realizada uma análise empírica baseada no tempo necessário para executar os procedimentos relativos ao treinamento da rede neural e sua posterior execução (obtenção da segmentação em cenas de um vídeo de entrada). O tempo necessário de execução foi estimado levando em conta as seguintes configurações de software e de hardware em um único computador individual.

- **Hardware**: Processador Intel Core i7 6700K, 64GB RAM DDR4 2400Mhz, placa de vídeo nVidia Geforce Titan V.

- **Software**: Sistema operacional Ubuntu 18.04, Torch7 na linguagem Lua (LuaTorch) aliada a uma biblioteca dedicada para redes recorrentes[1], além do framework CUDA 9.2.

[1] <https://github.com/Element-Research/rnn>

Na avaliação de desempenho das redes desenvolvidas, foi desconsiderado o tempo necessário para: 1) extrair e processar as características de baixo nível como MFCC, CSIFT, ConvFeat e BoW e; 2) carregar ou armazenar as redes neurais em tempo de treinamento ou segmentação. A avaliação do desempenho seguiu a metodologia previamente mencionado na Seção 5.2: uso do *leave-one-out* treinando sobre a BBC *Dataset*.

Utilizando a configuração supra-citada, o tempo necessário médio aproximado para o treinamento de um época, que consiste na execução e *back-propagation* de todos os vídeos da base de treinamento de entrada, é de 25s para a arquitetura baseada na fusão antecipada. Por outro lado, o tempo necessário para a arquitetura baseada na fusão tardia é de: 4s para a rede ConvFeat, 4s para a rede CSIFT, 16s para a rede MFCC e menos que 1s para a rede BoW. Além disso, o tempo necessário para a rede de fusão multimodal (arquitetura baseada na fusão tardia), é de menos que 1s. No total, cada época de treinamento da abordagem baseada na fusão tardia necessita de 22s de processamento. Graças a tais resultados, é possível concluir que a arquitetura baseada na fusão tardia (22s) obtém melhor desempenho em tempo de treinamento quando comparado ao tempo necessário para o treinamento de uma época da arquitetura baseada na fusão antecipada (25s). Resultado este devido ao maior número de pesos presente na rede neural baseada na fusão antecipada, que inclusive possui como entrada um maior volume de dados para cada amostra.

Em tempo de execução ou segmentação, o desempenho médio obtido pela arquitetura de fusão antecipada é muito superior a arquitetura de fusão tardia. Para um mesmo vídeo, o tempo necessário até obter a segmentação por meio da rede baseada na fusão tardia é de 2,31s (0,44s para ConvFeat, 1,08s para MFCC, 0,53s para CSIFT, 0,18s para BoW e 0,08s para a rede de fusão multimodal). A rede baseada na fusão antecipada requer apenas 1,35s para realizar a segmentação sobre as mesmas condições.

A rede baseada na fusão tardia pode ser adaptada de forma a melhorar seu desempenho por meio do uso de múltiplos computadores. No caso ideal, considerando máquinas idênticas, o processo de treinamento poderia ser realizado concorrentemente entre quatro computadores diferentes (um para cada característica utilizada) e, após o treinamento das redes unimodais, seguida pelo treinamento da rede de fusão multimodal. Em tal situação ideal, o tempo necessário para a execução de cada época seria de 17s (16s para a rede unimodal MFCC mais 1s relativo a rede de fusão multimodal), consideravelmente abaixo do tempo necessário pela rede baseada na fusão antecipada (que não suporta processamento concorrente de tal maneira). Em tempo de execução (segmentação), um processo idêntico poderia ser adotado, obtendo como resultado 1,24s para segmentar um vídeo de entrada, valor similar, mas também inferior ao necessário pela rede baseada na fusão antecipada.

É importante destacar que nenhum processo de otimização específico foi adotado com o intuito de otimizar o desempenho tanto do treinamento como da segmentação em cenas. Por exemplo, a aplicação do treinamento utilizando valores de pesos e gradientes com pontos flutu-

antes de 16bits (conhecido como FP16) resulta em ganhos expressivos quando combinado com a placa de vídeo no computador utilizado. A adoção de tal otimização poderia impactar negativamente a eficácia obtida pelas redes desenvolvidas, ou demandar um processo de treinamento específico e/ou mais complexo. Assim, para facilitar a reprodutibilidade do método proposto e por falta de suporte da biblioteca utilizada, tal otimização não foi adotada.

Por fim, é importante destacar que o tempo necessário para o processamento é diretamente dependente de fatores como o número de tomadas presente no vídeo e o número total de vídeos na base em tempo de treinamento. Outros fatores como a heterogeneidade da base de treinamento, como a presença de vídeos de diferentes domínios, tende a aumentar o número necessário de épocas para a convergência das redes neurais. Devido a definição de arquitetura e das características utilizadas, o método desenvolvido é independente do tamanho individual de cada tomada, haja visto que o número de vetores de características de entrada para cada característica utilizada é fixo. Por usar diretamente as características extraídas, o método é ainda independente de particularidades como resolução do vídeo, formato de codificação, entre outros.

5.4 Comparação com técnicas relacionadas

Para avaliar a eficácia da abordagem proposta perante o estado da arte reportado na literatura, foi realizada uma comparação dos resultados obtidos pelo modelo proposto com técnicas relacionadas, tanto unimodais como multimodais, baseadas ou não em Aprendizagem Profunda.

Para tal comparação, foram utilizadas as técnicas baseadas no alinhamento de sequências proposta por Chasanis, Likas e Galatsanos (2009) (NW), no grafo de transição de cenas multimodal proposta por Sidiropoulos *et al.* (2011) (STG) e nas redes neurais siamesas proposta por Baraldi, Grana e Cucchiara (2015a) (SDN), todas apresentadas nas Seção 3.2. Tais abordagens foram selecionadas seguindo os critérios adotados para considerar uma técnica como em estado da arte, conforme descrito no Capítulo 3.

Os valores obtidos por cada técnica supracitadas na BBC *Dataset* foram reportados no trabalho de Baraldi, Grana e Cucchiara (2015a), que, além da abordagem baseada em redes siamesas, propôs a própria BBC *Dataset*. Os resultados foram reportados exclusivamente na métrica F_{CO}, não constando os valores individuais de Cobertura e/ou Transbordamento.

É importante destacar que, conforme mencionado na Seção 5.2, a implementação da métrica de Cobertura, Transbordamento e F_{CO} foram providas por Baraldi, Grana e Cucchiara (2015a), garantindo a confiabilidade e imparcialidade do método de comparação empregado. Os resultados da técnica proposta, tanto usando a fusão antecipada como tardia, são reportados priorizando a métrica de F_{CO} (Tabelas 7 e 10, respectivamente). Os resultados da comparação entre a abordagem proposta, baseadas na fusão antecipada ou tardia, com as técnicas avaliadas reportadas no trabalho de Baraldi, Grana e Cucchiara (2015a) são apresentados na Tabela 12.

Tabela 12 – Comparação entre resultados obtidos pela abordagem proposta priorizando a métrica de F_{CO}, tanto usando a fusão antecipada ou tardia, com os resultados obtidos por técnicas relacionadas reportadas no trabalho de Baraldi, Grana e Cucchiara (2015a), na BBC *Dataset*, usando a métrica F_{CO}.

	Fusão antecipada	Fusão tardia	NW	STG	SDN
From Pole to Pole	61	**63**	39	47	56
Mountains	**72**	63	51	55	63
Ice Worlds	**68**	64	53	56	66
Great Plains	**68**	66	39	46	61
Jungles	**64**	60	56	42	55
Seasonal Forests	67	**73**	45	44	64
Fresh Water	62	62	53	**65**	59
Ocean Deep	**77**	69	45	63	64
Shallow Seas	**69**	59	36	61	64
Caves	**74**	64	19	58	64
Deserts	**78**	66	54	52	64
Média	**69**	64	45	54	62
Desvio padrão	5.54	4.02	10.92	7.94	**3.68**

Dentre as técnicas tradicionais, não baseadas em aprendizagem de máquina, a técnica multimodal STG proposta por Sidiropoulos *et al.* (2011) mostra uma clara vantagem sobre a técnica unimodal NW proposta por Chasanis, Likas e Galatsanos (2009), conforme esperado. A técnica SDN, por sua vez, baseada em aprendizagem de máquina, se mostrou mais eficaz tanto que NW e STG, sendo superado em apenas dois vídeos (**Jungles** e **Fresh Water**), conseguindo uma média 8% superior ao STG, mantendo um desvio padrão abaixo das demais (o menor valor entre todas as técnicas).

A técnica proposta baseada na fusão tardia obteve resultados significativos quando comparado as técnicas NW (a superou em todos os vídeos) e STG (a superou em todos os vídeos, com exceção de **Shallow Seas** e **Fresh Water**). Em relação a STG, a fusão tardia atingiu um maior valor médio de F_{CO} (64% para 54%) e menor desvio padrão (4.02 para 7.94) indicando, portanto, a obtenção de uma segmentação mais robusta e próxima à desejada.

Já em comparando a técnica de Aprendizagem Profunda baseada em redes siamesas (SDN), a rede baseada na fusão tardia proposta a superou em sete dos onze vídeos, com dois casos de eficácia idêntica (**Caves** e **Mountains**), obtendo um valor médio de F_{CO} 2% acima do obtido pela SDN. Mesmo no vídeo **Fresh Water**, no qual a STG apresenta uma eficácia acima de todas as abordagens inclusive as baseadas em Aprendizagem de Máquina, a rede adotada obtém resultados mais próximos do desejado e com pequena diferença nesse caso.

Já a abordagem baseada na fusão antecipada proposta obteve resultados claramente superiores ao reportado pela NW, STG e SDN, superando essa última, que obteve os melhores resultados médios reportados na literatura, em todos os vídeos analisados. Entre a abordagem de fusão antecipada e a SDN, a diferença obtida foi de 2% (**From Pole to Pole**) a até 14% (**Deserts**)

na métrica de F_{CO}. Em oito dos onze vídeos, a abordagem proposta obteve valores acima de 66%, o melhor valor obtido em qualquer vídeo pela SDN. Por fim, nos valores médios, a fusão antecipada obteve 69% de F_{CO}, 7% superior ao obtido pela SDN.

Em geral, é possível afirmar que tanto a abordagem baseada em fusão tardia como antecipada obtém resultados superiores ao reportados na literatura ao segmentar em cenas a BBC *Dataset*. Mesmo no único caso no qual uma técnica tradicional (não baseada em Aprendizagem Profunda) se sobressaiu as demais, a STG no vídeo **Fresh Water**, as redes propostas superam as demais técnicas, incluindo a SDN, por uma margem considerável.

Em termos de desempenho, as redes neurais utilizadas possuem custo computacional consideravelmente menor que a técnica baseada na rede siamesa (SDN). As demais técnicas não foram comparadas pois a) não são baseadas em abordagens de Aprendizagem de Máquina, b) utilizam outro conjunto de extratores de características com particularidades diversas ou c) utilizam apenas uma única modalidade. A SDN requer a construção de uma matriz de distâncias entre todas as tomadas, necessitando ao menos $\frac{n^2-n}{2}$ execuções da rede neural para a construção de tal matriz para um vídeo com n tomadas de duração. Após a matriz de distância ter sido calculada, o agrupamento espectral entre as tomadas é realizado, este também sendo um processo de considerável custo computacional. Por sua vez, a técnica proposta requer uma única execução da rede neural treinada para classificar todas as tomadas do vídeo, seguida de um algoritmo de baixo custo (Algoritmo 3).

5.5 Discussões sobre o capítulo

Neste capítulo foi descrita uma avaliação da eficácia da abordagem desenvolvida para a segmentação em cenas usando uma base de vídeos pública no domínio de documentários, chamada de BBC *Dataset*, previamente descrita na Subseção 2.6.1. Para a avaliação, as redes neurais desenvolvidas foram treinadas conforme processo descrito na Seção 5.1. A avaliação em si foi realizada por meio das principais métricas utilizadas na área, descritas na Subseção 2.6.2, seguindo a metodologia dos testes apresentada na Seção 5.2. Uma discussão em relação a eficácia e desempenho do método proposto ao realizar o processo de treinamento e de segmentação em cenas foi apresentada na Subseção 5.3.3. Por fim, uma comparação entre os resultados de eficácia obtidos pela abordagem proposta com outras técnicas de segmentação relacionadas, tanto unimodais como multimodais, tradicionais ou baseadas na aprendizagem de máquina, é descrito na Seção 5.4.

A eficácia foi medida utilizando as principais métricas adotadas por pesquisadores da área, tais como Precisão e Abrangência, métricas bem-conhecidas e utilizadas principalmente por trabalhos seminais, além da Cobertura e Transbordamento, criadas especificamente para a segmentação em cenas e amplamente adotadas em trabalhos recentes na área. Além disso, foram realizados testes com uma formulação alternativa da métrica de Transbordamento, chamada de

New Overflow (*NO*), melhor adaptada para determinados casos de sub-segmentação em trechos específicos de vídeo, no qual a formulação original do Transbordamento pode apresentar valores inválidos ou não-confiáveis.

Diferentes configurações de limiares, parâmetro essencial no algoritmo de identificação de cenas utilizado (Algoritmo 3), foram utilizadas, selecionados heuristicamente de maneira a priorizar as métricas de F_{PR}, F_{CO} ou F_{CNO}. Para evitar casos de sobre-especialização, as redes foram analisadas em diversas épocas diferentes, resultando em uma análise do melhor caso.

A arquitetura baseada na fusão antecipada, na qual todas as modalidades são analisadas conjuntamente para formar um vetor de característica único, mostrou uma maior variação de eficácia quando comparado com a rede baseada na fusão tardia. A fusão antecipada mostra uma eficácia acima da fusão tardia quando priorizada a mesma métrica específica. Tal fato indica que a fusão antecipada é capaz de, em dada configuração, melhor reconhecer tanto as tomadas de transição (medidas pela Precisão e Abrangência), como o agrupamento de tomadas nas cenas (medidas pela Cobertura e Transbordamento). Por sua vez, a fusão tardia demonstrou um equilíbrio da eficácia obtida nas diferentes configurações de limiar e de métrica a ser priorizada, resultando inclusive em menor valor de desvio padrão na F_{CO}.

Os resultados reportados por Baraldi, Grana e Cucchiara (2015a) foram utilizados para comparar as técnicas NW (CHASANIS; LIKAS; GALATSANOS, 2009), STG (SIDIROPOULOS *et al.*, 2011) e SDN (BARALDI; GRANA; CUCCHIARA, 2015a) com as redes propostas, usando as métricas de Cobertura e Transbordamento. O objetivo de tal comparação é o de validar a eficácia do método proposto frente a técnicas unimodais (NW), multimodais (STG) e baseadas na aprendizagem de máquina (SDN) em estado da arte reportadas na literatura. Ambas as arquiteturas desenvolvidas obtiveram resultados acima do obtido por NW e STG, com uma diferença média entre 15% (STG) a até 24% (NW) de F_{CO} com a arquitetura baseada na fusão antecipada. As arquiteturas propostas, usando tanto a fusão tardia como antecipada, superaram inclusive a SDN, tanto em valores médios como na maioria dos vídeos analisados. Quanto ao desempenho, a SDN necessita realizar uma série de execuções da rede neural siamesa para calcular o dissimilaridade tomada-a-tomada, aliada a um algoritmo caro de agrupamento espectral. Tais particularidades tornam a SDN consideravelmente custosa computacionalmente quando comparada as arquiteturas desenvolvidas, que necessitam de apenas uma única execução da rede neural, seguido de um algoritmo de identificação das transições de cenas de baixo custo, para segmentar um vídeo de tamanho qualquer.

Em termos de desempenho, a arquitetura do modelo proposto baseado na fusão tardia obteve melhores resultados quando comparado a arquitetura baseada na fusão tardia em tempo de treinamento. A soma do tempo de treinamento necessário para a convergência das quatro redes neurais unimodais (ConvFeat, CSIFT, MFCC e BoW) é inferior ao tempo necessário ao modelo baseado na fusão antecipada. Por outro lado, em tempo de execução (segmentação), a abordagem baseada na fusão antecipada apresenta leve vantagem sobre a fusão tardia, que necessita da

execução de quatro redes neurais até obter a segmentação final. Uma possível melhora do tempo de processamento da abordagem baseada em fusão tardia pode ser obtida, tanto em relação ao tempo de treinamento como de segmentação, por meio da adoção do processamento paralelo em diferentes computadores: por utilizar quatro redes neurais unimodais e independentes, é possível realizar o treinamento de cada rede unimodal de forma concorrente em diferentes dispositivos. Por fim, é importante ressalvar que, mesmo sem adotar tal abordagem, ambas as arquiteturas são capazes de segmentar trechos longos de vídeo em um tempo bastante reduzido

CAPÍTULO

6

CONCLUSÕES

Os objetivos deste trabalho foram alcançados. Os resultados reportados no Capítulo 5 demonstram uma vantagem significativa do método proposto sobre o estado da arte na tarefa de segmentação temporal de vídeo em cenas. Em média, as arquiteturas propostas, baseadas em diferentes abordagens de fusão multimodal, obtiveram de 2% (fusão tardia) a até 7% (fusão antecipada) de avanço sobre o estado da arte na métrica F_{CO} ao usar uma base de vídeos pública, a BBC *Dataset*, formada de vídeos no domínio de documentários. Adicionalmente, em apenas um único vídeo da referida base de vídeos o método proposto não obteve o melhor resultado da comparação realizada com outras técnicas em estado da arte usando diferentes abordagens (unimodal, multimodal e multimodal baseada em aprendizagem de máquina). Mesmo nesse caso, o método obteve o segundo melhor resultado reportado. Dentre as duas arquiteturas desenvolvidas, a avaliação indica resultados bastante similares, com pequena vantagem para a rede baseada em fusão antecipada quando considerada as métricas de Cobertura/Transbordamento. Tais resultados sugerem que a rede baseada em fusão antecipada é capaz de obter segmentações mais próximas da desejada devido a maior capacidade de correlacionar as características de entrada que a rede neural recorrente possui em tal configuração.

Além de superar os resultados de técnicas relacionadas em estado da arte, o segundo objetivo proposto também foi atingido, o de projetar um método de segmentação em cenas composto de duas partes: 1) uma modelagem do problema de segmentação em cenas como um problema de classificação de tomadas e 2) um modelo baseado em Aprendizagem Profunda para realizar as etapas de pré-processamento e de segmentação no processo de segmentação temporal de vídeo em cenas. O modelo proposto é formado pela combinação de redes convolucionais e recorrentes organizadas em duas arquiteturas diferentes, uma usando a abordagem de fusão tardia e a outra a abordagem de fusão antecipada. Tal modelo, parte integrante do método, oferece como benefício o suporte a características de diferentes modalidades (visual, aural e/ou textual) e particularidades para a tarefa de segmentação em cenas. Tal suporte auxilia a tornar o processo de segmentação em cenas independente das características de entrada, permitindo que diferentes

combinações de características possam ser avaliadas em busca de melhores resultados. Já a rede recorrente permite a análise do correlacionamento temporal entre as tomadas de entrada, auxiliando na identificação das tomadas pertencentes a uma mesma cena. Adicionalmente, o modelo pode ser instanciado, de maneira transparente, usando as abordagens de fusão tardia ou antecipada. Por fim, o modelo proposto é capaz de suportar diferentes algoritmos de identificação de transições de cenas e de pós-processamento, são fruto da independência entre o algoritmo de segmentação em cenas e a saída propriamente dita da referida arquitetura de rede neural utilizada.

Em termos de desempenho, a arquitetura baseada na fusão tardia apresenta significativa vantagem no tempo de treinamento de cada rede unimodal, como já esperado, requerendo um número menor de épocas de treinamento quando comparada com a arquitetura baseada na fusão antecipada. A soma do tempo necessário para o treinamento, por época, de cada rede neural unimodal e da rede de fusão multimodal é ainda inferior ao tempo necessário pela abordagem baseada na fusão antecipada. É importante destacar ainda que é possível otimizar o tempo de treinamento da abordagem baseada na fusão tardia por meio do treino das redes unimodais concorrentemente, em computadores diferentes. Com tal abordagem, a arquitetura de fusão tardia oferece um ganho expressivo de desempenho, superando a arquitetura baseada na fusão tardia tanto em tempo de treinamento como de segmentação. Nota-se que, independentemente da arquitetura sendo utilizada, o método proposto é capaz de segmentar um vídeo com horas de duração em poucos segundos.

É importante ressalvar, ainda, que o modelo de rede proposta pode ser facilmente estendida a outras tarefas de vídeo baseadas em classificação, tanto binárias como multi-classe. No primeiro caso, basta modificar a modelagem da tarefa anteriormente descrita de acordo com o novo problema. Já no segundo caso, além da modelagem, basta alterar a saída da rede neural recorrente, particularmente da função linear sigmoidal adotada, para outra função multi-classe adequada ao problema. Ao adotar uma função para problemas multi-classe, o processo de treinamento e especialmente a função de perda também deve ser adaptada.

As principais contribuições deste trabalho são: o método proposto em si, que consiste em uma abordagem de segmentação temporal de vídeo em cenas baseado em Aprendizagem Profunda; a modelagem proposta, responsável por modelar o problema da segmentação temporal de vídeo em cenas como um problema de classificação de tomadas; o modelo proposto, consistindo em uma abordagem de segmentação em cenas baseado em redes neurais convolucionais e recorrentes; suporte a diferentes arquiteturas do modelo, incluindo as abordagens de fusão antecipada e fusão tardia; a técnica de segmentação em cenas, baseada na comparação entre limiares, considerada de baixo custo.

Este trabalho deu origem, ainda, aos seguintes trabalhos publicados em eventos científicos:

- Barbieri, T. T. S.; Trojahn, T. H.; Ponti-Jr, M. P.; Goularte, R. **Shot-HR: A video shot representation method based on visual features**. ACM SAC, 2015.

- Kishi, R. M.; Trojahn, T. H.; Goularte. **An evaluation of readily usable automatic video shot segmentation techniques**. WebMedia, 2016.

- Kishi, R. M.; Trojahn, T. H.; Goularte. **Temporal video scene segmentation by fused bags-of-features**. WebMedia 2018.

- Trojahn, T. H.; Kishi, R. M.; Goularte, R. **A new multimodal deep-learning model to video scene segmentation**. WebMedia, 2018.

Dos artigos publicados em eventos científicos, o primeiro (**Shot-HR: A video shot representation method based on visual features**) têm por objetivo apresentar uma avaliação da eficácia de diferentes métodos de representação de tomadas, principalmente utilizando quadros-chave, na tarefa de segmentação em cenas. Já o segundo artigo mencionado (**An evaluation of readily usable automatic video shot segmentation techniques**) apresenta uma avaliação de técnicas de segmentação em tomadas cuja implementação esteja efetivamente disponível publicamente. O terceiro artigo, por sua vez, trata de uma abordagem multimodal de segmentação em cenas usando a análise do correlacionamento temporal entre diferentes características para detectar similaridade entre tomadas, que indicam tomadas pertencentes a mesma cena. Por fim, o artigo **A new multimodal deep-learning model to video scene segmentation** se refere a uma versão inicial deste trabalho, apresentando as abordagens de fusão antecipada e tardia sem usar a modalidade textual.

Foram publicados ainda dois artigos em diferentes periódicos: Lopes, B. L.; Trojahn, T. H. Goularte, R. **Video Scene Detection by Multimodal Bag of Features**. JIDM, 2014 e; Kishi, R. M.; Trojahn, T. H.; Goularte, R. **Correlation based feature fusion for the temporal video scene segmentation task**. MTAP, 2018. O primeiro consiste em uma abordagem de segmentação em cenas usando dicionários de palavras visuais (SIFT) e aurais (MFCC) usando a fusão multimodal tardia, obtendo resultados significativos em uma base de vídeos de filmes, com claro avanço sobre as abordagens unimodais individuais. O segundo, por sua vez, consiste na extensão do trabalho previamente mencionado e publicado em evento científico (**Temporal video scene segmentation by fused bags-of-features**), contendo melhorias nos testes de avaliação realizados e outros detalhes da implementação.

Durante o desenvolvimento deste trabalho, diferentes análises e considerações foram levantadas resultando em um conjunto de possíveis trabalhos futuros descritos a seguir:

- **Avaliação de eficácia em outras bases de vídeos:** neste trabalho, as redes desenvolvidas foram treinadas e aplicadas sobre a BBC *Dataset*, uma das bases de dados mais populares da área, formada de vídeos no domínio de documentários. Novas avaliações podem ser

realizadas sobre bases de dados diferentes de maneira a verificar a eficácia da técnica em tais condições. Tais testes podem ser realizados tanto com as redes treinadas na BBC *Dataset*, utilizadas no processo de avaliação realizada neste trabalho, como por meio de novo treinamento utilizando outro conjunto de dados de treinamento, com vídeos de domínios diversos.

- **Avaliar o impacto da flexibilidade na arquitetura baseada na fusão tardia:** uma das principais particularidades da fusão tardia, de fundir decisões obtidas de cada característica ou modalidade individual, permite um alto grau de flexibilidade pois facilita a alteração, seleção ou priorização de determinadas decisões individuais. Tal flexibilidade não foi explorada neste trabalho, podendo a mesma ser avaliada para casos específicos, como em vídeos cuja modalidade específica deve ter maior prioridade que outras modalidades.

- **Análise e tratamento do desbalanceamento da base de dados:** conforme mencionado, há um desbalanceamento significativo na base de dados BBC *Dataset* e outras potenciais bases de vídeos para a segmentação em cena, devido ao fato de haver um maior número de tomadas "não-transição" do que tomadas classificadas como "transição". O efeito de tal desbalanceamento não foi mensurado neste trabalho. A adoção de uma abordagem para balanceamento da base de dados, como a adoção de diferentes pesos para o erro médio detectado na função objetivo, pode resultar em melhorias tanto de eficácia, resultando em uma melhor segmentação, como também de desempenho, ao reduzir o tempo de treinamento graças a redução no número de épocas necessárias.

- **Avaliação da aplicação do modelo em outras tarefas relacionadas:** o modelo proposto, composto de redes convolucionais e recorrentes, embora aplicado a segmentação temporal hierárquica de vídeo em cenas, não é limitado a tal tarefa. Tal modelo pode ser aplicado a diversas tarefas de vídeo nas quais informações de tomadas anteriores possam ou devem ser consideradas, como a classificação de segmentos de vídeo.

- **Utilização de características de maior nível semântico:** a aplicação do modelo proposto na tarefa de segmentação em cenas utilizou características de baixo nível semântico, tal como vetores de característica CSIFT ou MFCC. Como o modelo provê suporte a adição de novas características e modalidades de entrada, o emprego de características com maior nível semântico, tais como as chamadas "características semânticas" propostas por Baraldi, Grana e Cucchiara (2017), pode resultar em melhores segmentações em cenas.

- **Análise e aprimoramento do procedimento de treinamento:** neste trabalho, o treinamento adotado é baseado no algoritmo de descida do gradiente estocástico com anelamento progressivo da taxa de aprendizagem. Embora amplamente utilizado em diversos trabalhos relacionados, a literatura reporta métodos recentes de otimização do treinamento, tais como o AdaGrad (DUCHI; HAZAN; SINGER, 2011), Adam (KINGMA; BA, 2014), Adadelta (ZEILER, 2012), AMSGrad (REDDI; KALE; KUMAR, 2018), entre outros (RUDER,

2016). Tal processo de treinamento pode resultar tanto em melhorias de desempenho, com menor tempo necessário para o treinamento, como também ganhos em eficácia.

REFERÊNCIAS

ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. **IEEE Transactions on Knowledge and Data Engineering**, IEEE Computer Society, Washington, DC, USA, v. 17, n. 6, p. 734–749, jun 2005. ISSN 1041-4347. Disponível em: <https://ieeexplore.ieee.org/document/1423975>. Citado na página 23.

ATREY, P. K.; HOSSAIN, M. A.; El Saddik, A.; KANKANHALLI, M. S. Multimodal fusion for multimedia analysis: a survey. **Multimedia Systems**, Springer-Verlag, v. 16, n. 6, p. 345–379, nov 2010. ISSN 1432-1882. Disponível em: <https://dx.doi.org/10.1007/s00530-010-0182-0>. Citado nas páginas 26, 47 e 48.

ATREY, P. K.; MADDAGE, N. C.; KANKANHALLI, M. S. Audio based event detection for multimedia surveillance. In: **IEEE International Conference on Acoustics Speech and Signal Processing Proceedings**. [s.n.], 2006. v. 5, p. 813–816. ISSN 1520-6149. Disponível em: <https://ieeexplore.ieee.org/document/1661400/>. Citado nas páginas 25 e 37.

BABER, J.; AFZULPURKAR, N.; BAKHTYAR, M. Video segmentation into scenes using entropy and SURF. In: **7th International Conference on Emerging Technologies**. Islamabad, Paquistan: IEEE, 2011. p. 1–6. ISBN 978-1-4577-0769-8. Disponível em: <https://ieeexplore.ieee.org/document/6048496/>. Citado nas páginas 24 e 33.

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: ACM, 2015. (MM '15), p. 1199–1202. ISBN 978-1-4503-3459-4. Disponível em: <https://dl.acm.org/citation.cfm?doid=2733373.2806316>. Citado nas páginas 17, 18, 25, 27, 33, 38, 39, 46, 51, 52, 65, 68, 69, 70, 71, 74, 81, 82, 84, 90, 94, 96, 105, 106 e 108.

____. Shot and scene detection via hierarchical clustering for re-using broadcast video. In: ____. **Computer Analysis of Images and Patterns: 16th International Conference on Computer Analysis of Images and Patterns**. Cham: Springer International Publishing, 2015. p. 801–811. ISBN 978-3-319-23192-1. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-319-23192-1_67>. Citado na página 54.

____. Recognizing and Presenting the Storytelling Video Structure with Deep Multimodal Networks. **IEEE Transactions on Multimedia**, v. 19, n. 5, p. 955–968, may 2017. ISSN 1520-9210. Disponível em: <https://ieeexplore.ieee.org/document/7797131/>. Citado nas páginas 25, 27, 33, 38, 71, 84, 86, 94 e 114.

BEAUFAYS, F.; ABDEL-MAGID, Y.; WIDROW, B. Application of neural networks to load-frequency control in power systems. **Neural Networks**, Elsevier Science Ltd., Oxford, UK, UK, v. 7, n. 1, p. 183–194, jan. 1994. ISSN 0893-6080. Disponível em: <https://dx.doi.org/10.1016/0893-6080(94)90067-1>. Citado na página 44.

BENGIO, Y.; SIMARD, P.; FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. **IEEE Transactions on Neural Networks**, IEEE Press, Piscataway, NJ,

USA, v. 5, n. 2, p. 157–166, mar 1994. ISSN 10459227. Disponível em: <https://ieeexplore.ieee.org/document/279181/>. Citado na página 44.

BLANKEN, H. M.; BLOK, H. E.; FENG, L.; VRIES, A. P. de (Ed.). **Multimedia Retrieval**. Springer, 2007. (Data-Centric Systems and Applications). ISBN 978-3-540-72894-8. Disponível em: <https://doi.org/10.1007/978-3-540-72895-5>. Citado nas páginas 24, 25, 32 e 34.

BOLLE, R. M.; YEO, B. L.; YEUNG, M. M. Video query: Research directions. **IBM Journal of Research and Development**, IBM, v. 42, n. 2, p. 233–252, mar 1998. ISSN 0018-8646. Disponível em: <https://ieeexplore.ieee.org/document/5389317/>. Citado na página 33.

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. Doi:10.5244/C.16.23. Citado na página 36.

BRUNELLI, R.; MICH, O.; MODENA, C. A survey on the automatic indexing of video data,. **Journal of Visual Communication and Image Representation**, v. 10, n. 2, p. 78–112, 1999. ISSN 1047-3203. Disponível em: <https://doi.org/10.1006/jvci.1997.0404>. Citado na página 24.

BURGHOUTS, G. J.; GEUSEBROEK, J. M. Performance evaluation of local colour invariants. **Computer Vision and Image Understanding**, Elsevier Inc., v. 113, n. 1, p. 48–62, 2009. ISSN 10773142. Disponível em: <https://dx.doi.org/10.1016/j.cviu.2008.07.003>. Citado na página 84.

CHASANIS, V.; KALOGERATOS, A.; LIKAS, A. Movie segmentation into scenes and chapters using locally weighted bag of visual words. In: **Proceedings of the ACM International Conference on Image and Video Retrieval**. New York, NY, USA: ACM, 2009. (CIVR '09), p. 35:1–35:7. ISBN 978-1-60558-480-5. Disponível em: <https://doi.acm.org/10.1145/1646396.1646439>. Citado nas páginas 35, 51, 63, 64, 65, 70, 71, 84, 88 e 89.

CHASANIS, V.; LIKAS, A.; GALATSANOS, N. Scene Detection in Videos Using Shot Clustering and Sequence Alignment. **IEEE Transactions on Multimedia**, IEEE, Washington, DC, USA, v. 11, n. 1, p. 89–100, jan 2009. ISSN 1520-9210. Disponível em: <https://ieeexplore.ieee.org/document/4721597/>. Citado nas páginas 14, 17, 24, 33, 63, 64, 65, 66, 67, 69, 70, 81, 82, 105, 106 e 108.

CHO, K.; MERRIËNBOER, B. van; GÜLÇEHRE, Ç.; BAHDANAU, D.; BOUGARES, F.; SCHWENK, H.; BENGIO, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: **Proceedings of the Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 2014. p. 1724–1734. Disponível em: <https://www.aclweb.org/anthology/D14-1179>. Citado na página 26.

CHUNG, J.; GULCEHRE, C.; CHO, K.; BENGIO, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: **NIPS 2014 Deep Learning and Representation Learning Workshop**. [S.l.: s.n.], 2014. Citado na página 44.

COIMBRA, D. B.; GOULARTE, R. Digital video scenes identification using audiovisual features. In: **Proceedings of the XV Brazilian Symposium on Multimedia and the Web**. New York, New York, USA: ACM Press, 2009. (WebMedia '09), p. 1–4. ISBN 9781605588803. Disponível em: <https://doi.acm.org/10.1145/1858477.1858520>. Citado na página 32.

DECHTER, R. Learning While Searching in Constraint-Satisfaction-Problems. In: **Proceedings of the 5th National Conference on Artificial Intelligence**. [s.n.], 1986. p. 178–185. Disponível em: <https://dl.acm.org/citation.cfm?id=2887799>. Citado na página 40.

DUCHI, J.; HAZAN, E.; SINGER, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. **The Journal of Machine Learning Research**, JMLR.org, v. 12, p. 2121–2159, 2011. ISSN 1532-4435. Disponível em: <https://dl.acm.org/citation.cfm?id=1953048.2021068>. Citado na página 114.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. 2. ed. New York, NY, USA: Wiley-Interscience, 2000. ISBN 0471056693. Citado na página 34.

EFRON, B. **The Jackknife, the Bootstrap, and Other Resampling Plans**. 3. ed. [S.l.]: Society for Industrial and Applied Mathematics, 1987. v. 1. (CBMS-NSF Regional Conference Series in Applied Mathematics, v. 1). ISBN 9780898711790. Citado na página 94.

ELMAN, J. L. Finding structure in time. **COGNITIVE SCIENCE**, Cognitive Science Society, v. 14, n. 2, p. 179–211, 1990. ISSN 0364-0213. Disponível em: <https://doi.org/10.1207/s15516709cog1402_1>. Citado nas páginas 26 e 44.

EVERINGHAM, M.; ESLAMI, S. M.; GOOL, L.; WILLIAMS, C. K.; WINN, J.; ZISSERMAN, A. The Pascal Visual Object Classes Challenge: A Retrospective. **International Journal of Computer Vision**, Kluwer Academic Publishers, Hingham, MA, USA, v. 111, n. 1, p. 98–136, 2015. ISSN 0920-5691. Disponível em: <https://dx.doi.org/10.1007/s11263-014-0733-5>. Citado na página 84.

FABRO, M. D.; BÖSZÖRMENYI, L. State-of-the-art and future challenges in video scene detection: a survey. **Multimedia Systems**, Springer Berlin Heidelberg, v. 19, n. 5, p. 427–454, oct 2013. ISSN 0942-4962. Disponível em: <https://ieeexplore.ieee.org/document/5972529/>. Citado nas páginas 24, 25, 26, 32, 50, 51, 54, 55 e 59.

FISCHLER, M. A.; BOLLES, R. C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. **Commun. ACM**, ACM, New York, NY, USA, v. 24, n. 6, p. 381–395, jun. 1981. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/358669.358692>. Citado na página 49.

FONSECA, M. S. **Combinando imagem e som para detecção de transições em vídeos digitais**. Dissertação (Mestrado) — Universidade Federal Fluminense, Niterói - RJ, Brasil, 2006. Disponível em: <http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=159539>. Citado na página 33.

GANCHEV, T.; FAKOTAKIS, N.; KOKKINAKIS, G. Comparative evaluation of various mfcc implementations on the speaker verification task. In: **in Proc. of the SPECOM-2005**. [S.l.: s.n.], 2005. p. 191–194. Citado na página 38.

GEUSEBROEK, J.-M.; BOOMGAARD, R. van den; SMEULDERS, A.; GEERTS, H. Color invariance. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 23, n. 12, p. 1338–1350, 2001. ISSN 01628828. Disponível em: <https://ieeexplore.ieee.org/document/977559/>. Citado na página 84.

GONZALEZ, R. C.; WOODS, R. C. **Processamento Digital de Imagens**. 3. ed. [S.l.]: Pearson Prentice Hall, 2009. ISBN 9788581435862. Citado nas páginas 34 e 41.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016. Disponível em: <http://www.deeplearningbook.org>. Citado nas páginas 27, 43, 44 e 83.

GOUYON, F.; PACHET, F.; DELERUE, O. On the use of zero-crossing rate for an application of classification of percussive sounds. In: **Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00**. [S.l.: s.n.], 2000. Citado na página 34.

GRAVES, A. Generating sequences with recurrent neural networks. **ArXiv e-print**, abs/1308.0850, 2013. Disponível em: <https://arxiv.org/abs/1308.0850>. Citado na página 45.

GRAVES, A.; JAITLY, N. Towards End-to-end Speech Recognition with Recurrent Neural Networks. In: **Proceedings of the 31st International Conference on International Conference on Machine Learning**. JMLR.org, 2014. (ICML'14), p. 1764–1772. Disponível em: <https://dl.acm.org/citation.cfm?id=3044805.3045089>. Citado na página 40.

GURNEY, K. **An Introduction to Neural Networks**. 1. ed. [S.l.]: CRC Press, 1997. Citado na página 40.

HAN, B.; WU, W. Video scene segmentation using a novel boundary evaluation criterion and dynamic programming. In: **IEEE International Conference on Multimedia and Expo**. [s.n.], 2011. p. 1–6. ISSN 1945-7871. Disponível em: <https://ieeexplore.ieee.org/document/6012001/>. Citado na página 56.

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**, IEEE, Washington, DC, USA, v. 9, n. 4, p. 580–588, jun 1999. ISSN 1051-8215. Disponível em: <https://ieeexplore.ieee.org/document/767124/>. Citado nas páginas 33, 54 e 96.

HARE, J. S.; SAMANGOOEI, S.; DUPPLAW, D. P. Openimaj and imageterrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In: **Proceedings of the 19th ACM international conference on Multimedia**. New York, NY, USA: ACM, 2011. (MM '11), p. 691–694. ISBN 978-1-4503-0616-4. Disponível em: <http://doi.acm.org/10.1145/2072298.2072421>. Citado na página 84.

HAYKIN, S. **Neural Networks and Learning Machines**. 3. ed. [S.l.]: Pearson Prentice Hall, 2009. ISBN 9780131471399. Citado na página 40.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: **Proceedings of the 2015 IEEE International Conference on Computer Vision**. Washington, DC, USA: IEEE Computer Society, 2015. (ICCV '15), p. 1026–1034. ISBN 978-1-4673-8391-2. Disponível em: <https://dx.doi.org/10.1109/ICCV.2015.123>. Citado nas páginas 26, 40, 42, 57 e 77.

____. Deep residual learning for image recognition. In: **IEEE Conference on Computer Vision and Pattern Recognition**. IEEE, 2016. (CVPR '16), p. 770–778. ISSN 1063-6919. Disponível em: <https://ieeexplore.ieee.org/document/7780459/>. Citado nas páginas 43, 76, 77 e 84.

Hildebrandt, M. Feature space fusion and feature selection for an enhanced robustness of the fingerprint forgery detection for printed artificial sweat. In: **2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)**. [S.l.: s.n.], 2015. p. 1–6. Citado na página 49.

HOCHREITER, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, World Scientific Publishing Co., Inc., River Edge, NJ, USA, v. 6, n. 2, p. 107–116, abr. 1998. ISSN 0218-4885. Disponível em: <http://dx.doi.org/10.1142/S0218488598000094>. Citado na página 42.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. **Neural Computation**, MIT Press, Cambridge, MA, USA, v. 9, n. 8, p. 1735–1780, nov 1997. ISSN 0899-7667. Disponível em: <https://dx.doi.org/10.1162/neco.1997.9.8.1735>. Citado na página 45.

HONG, Z. **SPEAKER GENDER RECOGNITION SYSTEM**. Dissertação (Mestrado) — University of Oulo, Oulo, Finlândia, 2017. Disponível em: <http://jultika.oulu.fi/files/nbnfioulu-201706082645.pdf>. Citado nas páginas 37 e 38.

HUANG, C.-R.; CHEN, C.-S.; CHUNG, P.-C. Contrast context histogram-an efficient discriminating local descriptor for object recognition and image matching. **Pattern Recognition**, Elsevier Science Inc., New York, NY, USA, v. 41, n. 10, p. 3071–3077, out. 2008. ISSN 0031-3203. Disponível em: <http://dx.doi.org/10.1016/j.patcog.2008.03.013>. Citado na página 63.

HUANG, F. J.; LECUN, Y. Large-scale Learning with SVM and Convolutional for Generic Object Categorization. In: **IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. IEEE, 2006. (CVPR, v. 1), p. 284–291. ISBN 0-7695-2597-0. Disponível em: <https://ieeexplore.ieee.org/document/1640771/>. Citado nas páginas 41 e 43.

Jesus, J.; Araújo, D.; Canuto, A. Fusion approaches of feature selection algorithms for classification problems. In: **2016 5th Brazilian Conference on Intelligent Systems (BRACIS)**. [S.l.: s.n.], 2016. p. 379–384. Citado nas páginas 25 e 49.

JIA, Y.; SHELHAMER, E.; DONAHUE, J.; KARAYEV, S.; LONG, J.; GIRSHICK, R.; GUADARRAMA, S.; DARRELL, T. Caffe: Convolutional architecture for fast feature embedding. In: **Proceedings of the 22Nd ACM International Conference on Multimedia**. New York, NY, USA: ACM, 2014. (MM '14), p. 675–678. ISBN 978-1-4503-3063-3. Disponível em: <http://doi.acm.org/10.1145/2647868.2654889>. Citado na página 68.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of Documentation**, v. 28, n. 1, p. 11–21, 1972. Disponível em: <https://doi.org/10.1108/eb026526>. Citado na página 39.

JOZEFOWICZ, R.; ZAREMBA, W.; SUTSKEVER, I. An empirical exploration of recurrent network architectures. In: **Proceedings of the 32nd International Conference on International Conference on Machine Learning**. JMLR.org, 2015. p. 2342–2350. Disponível em: <https://dl.acm.org/citation.cfm?id=3045367>. Citado nas páginas 44, 45, 46 e 77.

KINGMA, D. P.; BA, J. Adam: A Method for Stochastic Optimization. **International Conference on Learning Representations**, dec 2014. Disponível em: <https://arxiv.org/abs/1412.6980https://arxiv.org/abs/1412.6980>. Citado nas páginas 94 e 114.

KO, Y. A study of term weighting schemes using class information for text classification. In: **Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: ACM, 2012. (SIGIR '12), p. 1029–1030. ISBN 978-1-4503-1472-5. Disponível em: <https://doi.acm.org/10.1145/2348283.2348453>. Citado na página 39.

____. A new term-weighting scheme for text classification using the odds of positive and negative class probabilities. **Journal of the Association for Information Science and Technology**, Wiley-Blackwell, Hoboken, NJ, USA, v. 66, n. 12, p. 2553–2565, jan. 2015. ISSN 2330-1643. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23338>. Citado na página 39.

KOPRINSKA, I.; CARRATO, S. Temporal video segmentation: A survey. **Signal Processing: Image Communication**, v. 16, n. 5, p. 477–500, 2001. ISSN 0923-5965. Disponível em: <https://doi.org/10.1016/S0923-5965(00)00011-4>. Citado na página 32.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In: **Proceedings of the 25th International Conference on Neural Information Processing System**. Curran Associates Inc., 2012. (NIPS'12, v. 1), p. 1097–1105. Disponível em: <https://dl.acm.org/citation.cfm?id=2999134.2999257>. Citado nas páginas 42, 68, 76 e 77.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998. ISSN 00189219. Disponível em: <https://ieeexplore.ieee.org/document/726791/>. Citado na página 83.

LEWIS, D. D. **Representation and Learning in Information Retrieval**. Tese (Doutorado) — University of Massachusetts, Amherst, MA, USA, 1992. UMI Order No. GAX92-19460. Citado na página 39.

LI, X.; WU, X. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In: **2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. IEEE, 2015. p. 4520–4524. ISBN 978-1-4673-6997-8. ISSN 1520-6149. Disponível em: <https://ieeexplore.ieee.org/document/7178826/>. Citado na página 44.

LIANG, C.; ZHANG, Y.; CHENG, J.; XU, C.; LU, H. A novel role-based movie scene segmentation method. In: **Advances in Multimedia Information Processing**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. (PCM 2009, v. 5879 LNCS), p. 917–922. ISBN 978-3-642-10467-1. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-642-10467-1_82>. Citado nas páginas 24 e 33.

LIPTON, Z. C.; BERKOWITZ, J.; ELKAN, C. A critical review of recurrent neural networks for sequence learning. **Computing Research Repository - arXiv**, abs/1506.00019, 2015. Disponível em: <https://arxiv.org/abs/1506.00019>. Citado na página 26.

LOPES, B. L.; TROJAHN, T. H.; GOULARTE, R. Video Scene Detection by Multimodal Bag of Features. **Journal of Information and Data Management**, v. 5, n. 2, p. 194–205, jun 2014. ISSN 2178-7107. Disponível em: <https://seer.ufmg.br/index.php/jidm/article/view/632>. Citado nas páginas 24, 25, 35, 38, 46, 64, 65, 70, 71, 81, 82 e 84.

LOWE, D. G. Distinctive Image Features from Scale-Invariant Keypoints. **International Journal of Computer Vision**, v. 60, n. 2, p. 91–110, nov 2004. ISSN 0920-5691. Disponível em: <https://dl.acm.org/citation.cfm?id=993451.996342>. Citado nas páginas 35, 36, 37 e 84.

Lv, Y.; Zhou, W.; Tian, Q.; Sun, S.; Li, H. Retrieval oriented deep feature learning with complementary supervision mining. **IEEE Transactions on Image Processing**, v. 27, n. 10, p. 4945–4957, Oct 2018. ISSN 1057-7149. Citado na página 81.

MAAS, A. L.; HANNUN, A. Y.; NG, A. Y. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In: **Proceedings of the 30th International Conference on Machine Learning**. [s.n.], 2013. Disponível em: <https://pdfs.semanticscholar.org/367f/2c63a6f6a10b3b64b8729d601e69337ee3cc.pdf>. Citado na página 42.

MAJEED, S.; HUSAIN, H.; SAMAD, S.; IDBEAA, T. Mel frequency cepstral coefficients (MFCC) feature extraction enhancement in the application of speech recognition: A comparison study. **Journal of Theoretical and Applied Information Technology**, v. 79, p. 38–56, 09 2015. Citado na página 38.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. Scoring, term weighting, and the vector space model. In: ____. **Introduction to Information Retrieval**. [S.l.]: Cambridge University Press, 2008. p. 100–123. Citado na página 39.

MANZATO, M. G. **Uma arquitetura de personalização de conteúdo baseada em anotações do usuário**. Tese (Doutorado) — Universidade de São Paulo, São Carlos, SP, Brasil, 2 2011. An optional note. Disponível em: <https://dx.doi.org/10.11606/T.55.2011.tde-11042011-160836>. Citado na página 23.

MANZATO, M. G.; FORTES, R. P. M.; GOULARTE, R. **Técnicas e métodos para segmentacção de vídeo: um estudo sistemático**. [S.l.], 2006. 57 p. Disponível em: <https://conteudo.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_293.pdf>. Citado na página 59.

MARTINEZ, L. **Solução de problemas de otimização através de redes neurais multicamadas recorrentes**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos - SP, Brasil, 1996. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-29082017-101955/>. Citado na página 40.

MIKOLAJCZYK, K.; SCHMID, C. Performance evaluation of local descriptors. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE Computer Society, v. 27, n. 10, p. 1615–1630, oct 2005. ISSN 0162-8828. Disponível em: <https://dl.acm.org/citation.cfm?id=1083822.1083989>. Citado na página 84.

Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. **ArXiv e-prints**, jan. 2013. Disponível em: <https://arxiv.org/abs/1301.3781>. Citado nas páginas 39 e 69.

MÜLLER, M. **Information Retrieval for Music and Motion**. Berlin, Heidelberg: Springer-Verlag, 2007. ISSN 978-3-642-29166-1. ISBN 3540740473. Disponível em: <https://doi.org/10.1007/978-3-642-29166-1_27>. Citado na página 37.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: **Proceedings of the 27th International Conference on International Conference on Machine Learning**. Madison, WI, USA: Omnipress, 2010. (ICML'10), p. 807–814. ISBN 978-1-60558-907-7. Disponível em: <https://dl.acm.org/citation.cfm?id=3104322.3104425>. Citado nas páginas 42 e 77.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of Molecular Biology**, v. 48, n. 3, p. 443–453, 1970. ISSN 0022-2836. Disponível em: <https://www.sciencedirect.com/science/article/pii/0022283670900574>. Citado na página 66.

NG, A. Y.; JORDAN, M. I.; WEISS, Y. On Spectral Clustering: Analysis and an Algorithm. In: **Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic**. Cambridge, MA, USA: MIT Press, 2001. (NIPS'01), p. 849–856. Disponível em: <http://dl.acm.org/citation.cfm?id=2980539.2980649>. Citado na página 65.

NGUYEN, D. H.; WIDROW, B. Neural networks for self-learning control systems. **IEEE Control Systems Magazine**, v. 10, n. 3, p. 18–23, April 1990. ISSN 0272-1708. Citado na página 44.

NITANDA, N.; HASEYAMA, M.; KITAJIMA, H. Audio Signal Segmentation and Classification For Scene-Cut Detection. In: **IEEE International Symposium on Circuits and Systems**. IEEE, 2005. p. 4030–4033. ISBN 0-7803-8834-8. ISSN 02714310. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1465515>. Citado na página 67.

PASCANU, R.; MIKOLOV, T.; BENGIO, Y. On the difficulty of training recurrent neural networks. In: **Proceedings of the 30th International Conference on International Conference on Machine Learning**. JMLR.org, 2013. (ICML'13), p. 1310–1318. Disponível em: <https://dl.acm.org/citation.cfm?id=3042817.3043083>. Citado na página 44.

PETER, L.; MATEUS, D.; CHATELAIN, P.; SCHWORM, N.; STANGL, S.; MULTHOFF, G.; NAVAB, N. Leveraging random forests for interactive exploration of large histological images. In: GOLLAND, P.; HATA, N.; BARILLOT, C.; HORNEGGER, J.; HOWE, R. (Ed.). **Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014**. Cham: Springer International Publishing, 2014. p. 1–8. Citado na página 94.

PETERSOHN, C. Logical unit and scene detection: a comparative survey. In: GEVERS, T.; JAIN, R. C.; SANTINI, S. (Ed.). **Proc. SPIE 6820, Multimedia Content Access: Algorithms and Systems II**. [s.n.], 2008. v. 682002, n. January 2008, p. 682002–682002–17. ISBN 978-0-8194-6992-2. ISSN 0277-786X. Disponível em: <https://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=812768>. Citado na página 59.

PRASOON, A.; PETERSEN, K.; IGEL, C.; LAUZE, F.; DAM, E.; NIELSEN, M. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: MORI, K.; SAKUMA ICHIROAND SATO, Y.; BARILLOT, C.; NAVAB, N. (Ed.). **Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 246–253. ISBN 978-3-642-40763-5. Citado na página 41.

PROTASOV, S.; KHAN, A. M.; SOZYKIN, K.; AHMAD, M. Using deep features for video scene detection and annotation. **Signal, Image and Video Processing**, v. 12, n. 5, p. 991–999, jul 2018. ISSN 1863-1703. Disponível em: <https://doi.org/10.1007/s11760-018-1244-6>. Citado na página 24.

RASHEED, Z.; SHAH, M. Scene detection in Hollywood movies and TV shows. In: **2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.** Vancouver, Canada: IEEE Comput. Soc, 2003. v. 2, p. II–343–8. ISBN 0-7695-1900-8. Disponível em: <https://ieeexplore.ieee.org/document/1211489>. Citado nas páginas 33, 34, 51, 54, 61, 63, 64, 65, 70, 71, 88 e 89.

____. Detection and representation of scenes in videos. **IEEE Transactions on Multimedia**, IEEE, v. 7, n. 6, p. 1097–1105, Dec 2005. ISSN 1520-9210. Disponível em: <https://ieeexplore.ieee.org/document/1542086/>. Citado nas páginas 60 e 66.

RAZAVIAN, A. S.; AZIZPOUR, H.; SULLIVAN, J.; CARLSSON, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In: **IEEE Conference on Computer Vision and Pattern Recognition Workshops**. IEEE, 2014. p. 512–519. ISBN 978-1-4799-4308-1. ISSN 21607516. Disponível em: <http://ieeexplore.ieee.org/document/6910029/>. Citado nas páginas 81 e 83.

REDDI, S. J.; KALE, S.; KUMAR, S. On the Convergence of Adam and Beyond. In: **International Conference on Learning Representations**. [s.n.], 2018. Disponível em: <https://openreview.net/forum?id=ryQu7f-RZ>. Citado na página 114.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. **Revista de Sistemas de Informação da FSMA**, v. 7, p. 7–21, 2011. ISSN 19835604. Disponível em: <http://www.fsma.edu.br/si/7edicao.html>. Citado na página 39.

RICHARDSON, I. E. **Video Codec Design: Developing Image and Video Compression Systems**. New York, NY, USA: John Wiley & Sons, Inc., 2002. ISBN 0471485535. Citado na página 32.

RIJSBERGEN, C. G. **Information Retrieval**. 2. ed. London: Butterworths, 1979. 224 p. Citado na página 53.

ROBBINS, H.; MONRO, S. A Stochastic Approximation Method. **Ann. Math. Statist.**, The Institute of Mathematical Statistics, v. 22, n. 3, p. 400–407, 1951. Disponível em: <https://doi.org/10.1214/aoms/1177729586>. Citado na página 95.

ROTMAN, D.; PORAT, D.; ASHOUR, G. Robust and efficient video scene detection using optimal sequential grouping. In: **2016 IEEE International Symposium on Multimedia (ISM)**. [s.n.], 2016. p. 275–280. Disponível em: <https://ieeexplore.ieee.org/document/7823628>. Citado nas páginas 51 e 52.

RUDER, S. An overview of gradient descent optimization algorithms. **Computing Research Repository - arXiv**, abs/1609.04747, 2016. Disponível em: <https://arxiv.org/abs/1609.04747>. Citado na página 115.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. In: ANDERSON, J. A.; ROSENFELD, E. (Ed.). **Neurocomputing: Foundations of Research**. Cambridge, MA, USA: MIT Press, 1988. p. 696–699. ISBN 0-262-01097-6. Disponível em: <https://dl.acm.org/citation.cfm?id=65669.104451>. Citado nas páginas 26 e 44.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M.; BERG, A. C.; FEI-FEI, L. ImageNet Large Scale Visual Recognition Challenge. **International Journal of Computer Vision**, Springer US, v. 115, n. 3, p. 211–252, dec 2015. ISSN 1573-1405. Disponível em: <https://dx.doi.org/10.1007/s11263-015-0816-y>. Citado na página 40.

SACHDEVA, V. D.; BABER, J.; BAKHTYAR, M.; ULLAH, I.; NOOR, W.; BASIT, A. Performance evaluation of sift and convolutional neural network for image retrieval. **International Journal of Advanced Computer Science and Applications**, The Science and Information Organization, v. 8, n. 12, 2017. Disponível em: <http://dx.doi.org/10.14569/IJACSA.2017.081268>. Citado nas páginas 41 e 43.

SAHIDULLAH, M.; SAHA, G. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. **Speech Communication**, v. 54, n. 4, p. 543–565, 2012. ISSN 0167-6393. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167639311001622>. Citado nas páginas 37 e 38.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information Processing and Management: an International Journal**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 24, n. 5, p. 513–523, ago. 1988. ISSN 0306-4573. Disponível em: <https://dx.doi.org/10.1016/0306-4573(88)90021-0>. Citado na página 39.

SAWAI, K.; TAKAHASHI, T.; DEGUCHI, D.; IDE, I.; MURASE, H. Scene segmentation of wedding party videos by scenario-based matching with example videos. In: **Proceedings of the 19th ACM international conference on Multimedia**. New York, New York, USA: ACM Press, 2011. (MM '11), p. 1545. ISBN 9781450306164. Disponível em: <https://dl.acm.org/citation.cfm?id=2072298.2072061>. Citado na página 33.

SCHROFF, F.; KALENICHENKO, D.; PHILBIN, J. Facenet: A unified embedding for face recognition and clustering. In: **IEEE Conference on Computer Vision and Pattern Recognition**. 2015, 2015. p. 815–823. ISSN 1063-6919. Disponível em: <https://ieeexplore.ieee.org/document/7298682/>. Citado na página 26.

SHEN, Y.; DEMARTY, C.-H.; DUONG, N. Q. K. Deep learning for multimodal-based video interestingness prediction. In: **IEEE International Conference on Multimedia and Expo**. IEEE, 2017. (ICME 2017, July), p. 1003–1008. ISBN 978-1-5090-6067-2. ISSN 1945788X. Disponível em: <http://ieeexplore.ieee.org/document/8019300/>. Citado nas páginas 25 e 84.

SHI, J.; MALIK, J. Normalized cuts and image segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 8, p. 888–905, Aug 2000. ISSN 0162-8828. Disponível em: <https://ieeexplore.ieee.org/document/868688>. Citado na página 61.

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**, IEEE, v. 21, n. 8, p. 1163–1177, Aug 2011. ISSN 1051-8215. Disponível em: <https://ieeexplore.ieee.org/document/5742987/>. Citado nas páginas 17, 25, 33, 38, 46, 51, 61, 65, 66, 67, 69, 70, 71, 81, 82, 105, 106 e 108.

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**, IEEE, Piscataway, NJ, USA, v. 22, n. 6, p. 904–914, June 2012. ISSN 1051-8215. Disponível em: <https://ieeexplore.ieee.org/document/6111460/>. Citado nas páginas 24 e 57.

SIMO-SERRA, E.; TRULLS, E.; FERRAZ, L.; KOKKINOS, I.; FUA, P.; MORENO-NOGUER, F. Discriminative learning of deep convolutional feature point descriptors. In: **IEEE International Conference on Computer Vision**. IEEE, 2015. p. 118–126. ISSN 2380-7504. Disponível em: <https://ieeexplore.ieee.org/document/7410379/>. Citado na página 83.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **ArXiv e-prints**, abs/1409.1556, 2014. Disponível em: <https://arxiv.org/abs/1409.1556>. Citado nas páginas 41, 43, 76, 77 e 84.

SMEATON, A. F.; OVER, P.; DOHERTY, A. R. Video shot boundary detection: Seven years of TRECVid activity. **Computer Vision and Image Understanding**, Elsevier Inc., v. 114, n. 4, p. 411–418, apr 2010. ISSN 10773142. Disponível em: <https://dx.doi.org/10.1016/j.cviu.2009.03.011https://linkinghub.elsevier.com/retrieve/pii/S1077314209000587>. Citado na página 24.

SMEULDERS, A.; 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**, IEEE Computer Society, Washington, DC, USA, v. 22, n. 12, p. 1349–1380, 2000. ISSN 01628828. Disponível em: <https://dl.acm.org/citation.cfm?id=357871.357873>. Citado nas páginas 24 e 35.

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**. New York, New York, USA: ACM Press, 2005. p. 399–402. ISBN 1595930442. Disponível em: <https://portal.acm.org/citation.cfm?doid=1101149.1101236>. Citado nas páginas 26 e 48.

STEVENS, S. S.; VOLKMANN, J.; NEWMAN, E. B. A Scale for the Measurement of the Psychological Magnitude Pitch. **The Journal of the Acoustical Society of America**, v. 8, n. 3, p. 185–190, jan 1937. ISSN 0001-4966. Disponível em: <http://asa.scitation.org/doi/10.1121/1.1915893>. Citado na página 37.

TAN, C.-M.; WANG, Y.-F.; LEE, C.-D. The use of bigrams to enhance text categorization. **Information Processing and Management: an International Journal**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 38, n. 4, p. 529–546, jul. 2002. ISSN 0306-4573. Disponível em: <http://dx.doi.org/10.1016/S0306-4573(01)00045-0>. Citado na página 39.

TAN, T.; QIAN, Y.; YU, D.; KUNDU, S.; LU, L.; SIM, K. C.; XIAO, X.; ZHANG, Y. Speaker-aware training of LSTM-RNNS for acoustic modelling. In: **2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. IEEE, 2016. v. 2016-May, n. 61222208, p. 5280–5284. ISBN 978-1-4799-9988-0. ISSN 15206149. Disponível em: <https://ieeexplore.ieee.org/document/7472685/>. Citado na página 44.

TOFFLER, A. **Future Shock**. 1. ed. New York, NY, USA: Bantam, 1984. 576 p. Citado na página 23.

TROJAHN, T. H.; GOULARTE, R. Video scene segmentation by improved visual shot coherence. In: **Proceedings of the 19th Brazilian Symposium on Multimedia and the Web**. New York, NY, USA: ACM, 2013. (WebMedia '13), p. 23–30. ISBN 978-1-4503-2559-2. Disponível em: <https://doi.acm.org/10.1145/2526188.2526206>. Citado nas páginas 63, 70, 71, 82, 88 e 89.

TSUNOO, E.; BELL, P.; RENALS, S. Hierarchical Recurrent Neural Network for Story Segmentation. In: **Interspeech 2017**. ISCA: ISCA, 2017. p. 2919–2923. Disponível em: <https://dx.doi.org/10.21437/Interspeech.2017-392>. Citado na página 44.

TSUNOO, E.; KLEJCH, O.; BELL, P.; RENALS, S. Hierarchical recurrent neural network for story segmentation using fusion of lexical and acoustic features. In: **IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)**. IEEE, 2017. p. 525–532. ISBN 978-1-5090-4788-8. Disponível em: <https://ieeexplore.ieee.org/document/8268981/>. Citado na página 44.

UYSAL, A. K.; GUNAL, S. The impact of preprocessing on text classification. **Information Processing and Management: an International Journal**, Pergamon Press, Inc., Tarrytown, NY, USA, v. 50, n. 1, p. 104–112, jan. 2014. ISSN 0306-4573. Disponível em: <https://dx.doi.org/10.1016/j.ipm.2013.08.006>. Citado na página 39.

VENDRIG, J.; WORRING, M. Systematic evaluation of logical story unit segmentation. **IEEE Transactions on Multimedia**, Piscataway, NJ, USA, v. 4, n. 4, p. 492–499, dez. 2002. ISSN 1520-9210. Disponível em: <https://ieeexplore.ieee.org/document/1176947/>. Citado nas páginas 54 e 59.

VINCIARELLI, A.; FAVRE, S. Broadcast news story segmentation using social network analysis and hidden markov models. **Proceedings of the 15th international conference on Multimedia - MULTIMEDIA '07**, ACM Press, New York, New York, USA, p. 261, 2007. Disponível em: <https://portal.acm.org/citation.cfm?doid=1291233.1291287>. Citado na página 57.

VUKOTIC, V.; RAYMOND, C.; GRAVIER, G. A Crossmodal Approach to Multimodal Fusion in Video Hyperlinking. **IEEE MultiMedia**, v. 25, n. 2, p. 11–23, 2018. ISSN 1070-986X. Disponível em: <https://ieeexplore.ieee.org/document/8424826/>. Citado nas páginas 25 e 84.

WANG, J.; DUAN, L.; LU, H.; JIN, J.; XU, C. A Mid-Level Scene Change Representation Via Audiovisual Alignment. In: **IEEE International Conference on Acoustics Speed and Signal Processing Proceedings**. IEEE, 2006. v. 2, p. 409–412. ISBN 1-4244-0469-X. Disponível em: <https://ieeexplore.ieee.org/document/1660366/>. Citado nas páginas 24 e 33.

WERBOS, P. J. Backpropagation through time: what it does and how to do it. **Proceedings of the IEEE**, v. 78, n. 10, p. 1550–1560, Oct 1990. ISSN 0018-9219. Citado na página 44.

WIATOWSKI, T.; BOLCSKEI, H. A Mathematical Theory of Deep Convolutional Neural Networks for Feature Extraction. **IEEE Transactions on Information Theory**, v. 64, n. 3, p. 1845–1866, mar 2018. ISSN 0018-9448. Disponível em: <https://ieeexplore.ieee.org/document/8116648/>. Citado nas páginas 26 e 41.

WILSON, K.; DIVAKARAN, A. Discriminative genre-independent audio-visual scene change detection. In: **SPIE Conference on Multimedia Content Access: Algorithms and Systems**. [s.n.], 2009. v. 7255. Disponível em: <http://www.merl.com/publications/TR2009-001>. Citado na página 67.

WU, P.; MANJUNANTH, B. S.; NEWSAM, S. D.; SHIN, H. D. A texture descriptor for image retrieval and browsing. In: **Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries**. Fort Collins, CO, USA, USA: IEEE, 1999. (CBAIVL '99), p. 3–7. ISSN 0-7695-0034-X. Disponível em: <https://ieeexplore.ieee.org/document/781114/>. Citado na página 35.

WU, S.; JIN, M. Study on a New Video Scene Segmentation Algorithm. **Applied Mathematics & Information Sciences**, Natural Sciences Publishing, v. 9, n. 1, p. 361–368, 2015. ISSN 2325-0399. Disponível em: <https://dx.doi.org/10.12785/amis/090142>. Citado na página 38.

WU, Z.; JIANG, Y.-G.; WANG, X.; YE, H.; XUE, X. Multi-Stream Multi-Class Fusion of Deep Networks for Video Classification. In: **Proceedings of the 2016 ACM on Multimedia Conference**. New York, NY, USA: ACM Press, 2016. (MM '16), p. 791–800. ISBN 9781450336031. Disponível em: <https://doi.acm.org/10.1145/2964284.2964328>. Citado nas páginas 40 e 43.

XU, B.; WANG, N.; CHEN, T.; LI, M. Empirical Evaluation of Rectified Activations in Convolutional Network. **eprint arXiv**, may 2015. Disponível em: <https://arxiv.org/abs/1505.00853>. Citado nas páginas 42 e 77.

XU, K.; XIE, L.; YAO, K. Investigating LSTM for punctuation prediction. In: **2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)**. IEEE, 2016. p. 1–5. ISBN 978-1-5090-4294-4. Disponível em: <https://ieeexplore.ieee.org/document/7918492/>. Citado na página 44.

YAN, K.; WANG, Y.; LIANG, D.; HUANG, T.; TIAN, Y. Cnn vs. sift for image retrieval: Alternative or complementary? In: **Proceedings of the 24th ACM International Conference on Multimedia**. New York, NY, USA: ACM, 2016. (MM '16), p. 407–411. ISBN 978-1-4503-3603-1. Disponível em: <http://doi.acm.org/10.1145/2964284.2967252>. Citado na página 81.

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, jul 1998. ISSN 10773142. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1077314297906287>. Citado nas páginas 60, 61, 64, 66 e 71.

YU, J.; XIAO, X.; XIE, L.; CHNG, E. S.; LI, H. A DNN-HMM Approach to Story Segmentation. In: **Interspeech 2016**. [s.n.], 2016. p. 1527–1531. Disponível em: <https://dx.doi.org/10.21437/Interspeech.2016-873>. Citado na página 44.

YU, J.; XIE, L.; XIAO, X.; CHNG, E. S. A hybrid neural network hidden Markov model approach for automatic story segmentation. **Journal of Ambient Intelligence and Humanized Computing**, Springer Berlin Heidelberg, v. 8, n. 6, p. 925–936, nov 2017. ISSN 1868-5137. Disponível em: <https://link.springer.com/10.1007/s12652-017-0501-9>. Citado na página 44.

ZEILER, M. D. ADADELTA: An Adaptive Learning Rate Method. **arXiv e-Prints**, dec 2012. ISSN 09252312. Disponível em: <https://arxiv.org/abs/1212.5701>. Citado na página 114.


ICMC
USP
SÃO CARLOS