UNIVERSIDADE ESTADUAL PAULISTA
“Júlio de Mesquita Filho”

Pós-Graduação em Ciência da Computação

Lucas Barbosa de Almeida

Aprendizado Não Supervisionado para Recuperação Multimídia Multimodal

Rio Claro – SP

2022

Lucas Barbosa de Almeida

Aprendizado Não Supervisionado para Recuperação Multimídia Multimodal

Orientador: Prof. Dr. Daniel Carlos Guimarães Pedronette

Dissertação de Mestrado elaborada junto ao Programa de Pós-Graduação em Ciência da Computação – Área de Concentração em Inteligência Computacional como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Rio Claro - SP

2022

A447a

Almeida, Lucas Barbosa de
Aprendizado Não Supervisionado para Recuperação Multimídia Multimodal / Lucas Barbosa de Almeida. -- Rio Claro, 2022
86 p. : il.

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Instituto de Geociências e Ciências Exatas, Rio Claro
Orientador: Daniel Carlos Guimarães Pedronette

1. Aprendizado Não Supervisionado. 2. Recuperação Multimídia Multimodal por Conteúdo. 3. Redes Convolucionais Baseadas em Grafo. 4. Deep Graph Infomax. 5. Aprendizado de Representações. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Geociências e Ciências Exatas, Rio Claro. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Lucas Barbosa de Almeida

Aprendizado Não Supervisionado para Recuperação Multimídia Multimodal

Dissertação de Mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista “Júlio de Mesquita Filho”, como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação

Comissão Examinadora

Prof. Dr. Daniel Carlos Guimarães Pedronette
IGCE / UNESP/Rio Claro (SP)

Prof. Dr. Rodrigo Tripodi Calumby
UEFS / Feira de Santana (BA)

Prof. Dr. Fabricio Aparecido Breve
IGCE / UNESP/Rio Claro (SP)

Conceito: Aprovado.

Rio Claro/SP,
30 de Março de 2022

Agradecimentos

Agradeço ao apoio da **Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)** pela concessão da bolsa de pesquisa no processo nº **2020/03311-0**. As opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade dos autores e não necessariamente refletem a visão da FAPESP.

Ao fim dessa caminhada do mestrado, desses dois anos de muito esforço e empenho gostaria de agradecer a algumas pessoas que me foram fundamentais e me ajudaram na realização desse sonho, por isso expresso minha sincera gratidão a todas elas. Agradeço a minha irmã, provavelmente a pessoa mais determinada que eu já conheci, por sempre me inspirar a ser uma pessoa melhor e sempre me encorajar e ajudar na minha caminhada. Também agradeço imensamente ao meu orientador, professor Daniel, por toda a paciência, confiança e apoio desde o momento em que no meu terceiro ano de graduação resolvi desenvolver um projeto de Iniciação Científica, por ter me ensinado, orientado e motivado desde então. Sou grato ao meu amigo Victor, por quase uma década que me apoia e me motiva. Agradeço também às minhas amigas Alline e Soraya que estiveram presentes e me apoiaram em grande parte dos momentos mais difíceis dessa caminhada e que apesar de eu conhecer a relativamente pouco tempo se tornaram importantes para mim. Também sou imensuravelmente grato aos meus pais, pois todo o apoio deles me permitiu chegar onde cheguei. Por fim, agradeço a todo o corpo docente do programa de pós-graduação.

Resumo

Dado o crescimento vertiginoso de coleções multimídia, sejam vídeos, áudios ou imagens e a carência de dados rotulados, torna-se fundamental investigar abordagens não supervisionadas de recuperação de informação baseada no conteúdo. Considerando que informações de diferentes modalidades ou representações de um mesmo objeto tendem a ser complementares, é imprescindível explorar múltiplas modalidades no processo de recuperação de informação. Contudo, ao utilizar informações de modalidades distintas, depara-se com o desafio de como combinar as informações dessas diferentes fontes. No contexto dessa dissertação, serão investigadas abordagens de combinação utilizando múltiplos ranqueamentos por meio de métodos de aprendizado não supervisionado. De modo geral, tais métodos exploram relações contextuais entre os objetos, geralmente codificadas nas informações de similaridade das coleções, sem requerer dados rotulados ou intervenção de usuários. Além disso, foram consideradas abordagens recentes de redes convolucionais baseadas em grafos (*Graph Convolutional Networks* - GCNs). O treinamento de GCNs é tradicionalmente realizado de modo que cada nó se comunica com sua vizinhança, incorporando a si informações dos nós aos quais apresenta conexões no grafo. Neste trabalho, combinamos a capacidade de métodos de aprendizado não supervisionado em explorar a geometria do conjunto de dados e definir uma medida contextual de distância com a capacidade de GCNs de criar uma representação mais eficaz de cada instância para aprimorar os resultados de recuperação de vídeos em cenários não supervisionados e multimodais. Deste modo, o trabalho apresenta um levantamento bibliográfico, discute métodos de extração de características em diferentes modalidades, e apresenta propostas de métodos para recuperação multimídia capazes de combinar as informações de diferentes modalidades em dois cenários distintos. No primeiro cenário, são propostas diferentes abordagens para recuperação de vídeos considerando informações de diferentes modalidades (imagens, áudios e vídeos) e utilizando técnicas de aprendizado não supervisionado baseadas em ranqueamento e GCNs treinadas de modo não supervisionado. No segundo cenário, é proposto um método de aprendizado de representações para recuperação de imagens baseado na fusão de representações multimodais. A representação de cada imagem é obtida através de características extraídas de uma sequência composta de sua k-vizinhança mais próxima, também utilizando técnicas de aprendizado não supervisionado.

Palavras-chave: Aprendizado Não Supervisionado, Recuperação Multimídia por Conteúdo, Recuperação Multimodal de Informações, Aprendizado de Representações, Redes Convolucionais Baseadas em Grafos, *Deep Graph Infomax*.

Abstract

Given the rapid growth of multimedia collections, whether videos, audios or images, and the lack of labeled data, it is essential to investigate unsupervised approaches to content-based information retrieval. Considering that information from different modalities or representations of the same object tend to be complementary, it is essential to explore multiple modalities in the information retrieval process. However, when using information from different modalities, one is faced with the challenge of how to combine information from these different sources. In the context of this dissertation, combination approaches using multiple rankings through unsupervised learning methods will be investigated. In general, such methods explore contextual relationships between objects, usually encoded in the similarity information of the collections, without requiring labeled data or user intervention. Furthermore, recent approaches to graph-based convolutional networks (*Graph Convolutional Networks* - GCNs) were considered. The training of GCNs is traditionally performed so that each node communicates with its neighborhood, incorporating information from the nodes to which it has connections in the graph. In this work, we combine the ability of unsupervised learning methods to explore the geometry of the dataset and define a contextual measure of distance with the ability of GCNs to create a more effective representation of each instance to improve video retrieval results in unsupervised and multimodal scenarios. In this way, the work presents a bibliographic survey, discusses methods for extracting features in different modalities, and presents proposals for methods for multimedia retrieval capable of combining information from different modalities in two different scenarios. In the first scenario, different approaches are proposed for video retrieval considering information from different modalities (images, audios and videos) and using unsupervised learning techniques based on ranking and unsupervised trained GCNs. In the second scenario, a representation learning method for image retrieval based on the fusion of multimodal representations is proposed. The representation of each image is obtained through features extracted from a sequence composed of its nearest k-neighborhood, also using unsupervised learning techniques.

Keywords: Unsupervised Learning, Multimedia Content Retrieval, Multimodal Information Retrieval, Representation Learning, Graph Convolutional Networks, Deep Graph Infomax.

Lista de ilustrações

Lista de tabelas

Símbolos

$\mathcal{X}$	Coleção de Objetos
f	Função de Extração de Características
$\mathbf{v}$	Vetor de Características
ρ	Função de Distância
τ	Lista Ranqueada
$\mathcal{T}$	Conjunto de Listas Ranqueadas
g_S	Função que Obtém o Ranqueamento de Melhor Eficácia.
D_x	Conjunto finito de pontos (pixels) em $\mathbb{N}^2$
I_x	Função que atribui a cada pixel $p \in D_x$ um vetor $I(p) \in \mathbb{N}^3$ (uma cor no sistema RGB é atribuída a um pixel).
σ	Sequência de imagens (ou quadros)
$N_2(x_q, k_r)$	Conjunto de k_r mais semelhante a x_q de acordo com as características extraídas pela CNN 2D e listas ranqueadas
S	Conjunto de Sequências
$\mathbf{A}$	Matriz de Adjacência
$\mathbf{X}$	Matriz de *Features*
f_{gcn}	Função que gera *embenddings* a partir de dados de entrada, utilizando uma GCN
G	Grafo $G(X, A)$, que possui $\mathbf{X}$ como conjunto de arestas e $\mathbf{A}$ como conjunto de vértices
$\mathbf{Z}$	Matriz de *embenddings* gerada por uma GCN
E	*Enconder* obtido por meio do algoritmo DGI, tal que $\mathcal{E} : \mathbb{R}^{N \times F} \times \mathbb{R}^{N \times N} \rightarrow \mathbb{R}^{N \times F'}$

Sumário

1 Introdução

Dado o aumento vertiginoso no volume e diversidade de coleções multimídia, com dados não rotulados a expandir-se diariamente e incluindo vídeos, áudios e imagens (DATTA et al., 2008), torna-se premente o desenvolvimento de estratégias capazes de analisar e recuperar mídias baseado no conteúdo. Destaca-se que rotular manualmente todas essas coleções seria algo inviável e de custo proibitivo, em especial pela quantidade de instâncias mas também pela subjetividade do conteúdo presente nessas mídias, torna-se então necessário investigar abordagens automatizadas. Todavia, automatizar tal processo é uma tarefa não trivial e desafiadora, em especial se considerada a multimodalidade inerente aos dados disponíveis em conteúdo multimídia.

A área de recuperação de informação compreende diversos desafios, inclusive os relacionados à sub-área de recuperação de informação pelo conteúdo, com a qual este projeto está relacionado. A área de recuperação de informação iniciou considerando buscas baseadas em conteúdo textual e tem evoluindo de maneira acelerada nos últimos anos, oferecendo atualmente resultados notáveis em grande parte dos cenários existentes. Contudo, apesar da eficácia e desempenho dos sistemas de recuperação textual terem atingido um elevado grau de maturidade, ainda há desafios na área de recuperação, especialmente em aplicações de recuperação de conteúdos multimídia, devido a riqueza de conteúdo e subjetividade na interpretação dos dados (KOFLER; LARSON; HANJALIC, 2016). Dessa forma, a fim de explorar o conteúdo intrínseco das coleções de dados e amenizar o impacto das dificuldades citadas, os sistemas de recuperação por conteúdo têm sido propostos em diversas abordagens para diferentes domínios, como os sistemas de recuperação de imagem para diagnóstico de doenças (AGARWAL; MOSTAFA, 2011), reconhecimento facial (SULTANA; GAVRILOVA, 2013), sensoriamento remoto (WANG; SONG, 2013), identificação de objetos (SCHOBER; HERMES; HERZOG, 2004). Considerando outras modalidades, também há utilizações em diversos domínios, como os sistemas de áudio aplicados em reconhecimento de gênero de músicas, padrões (Tzanetakis; Cook, 2002) e reconhecimento de discurso (Riccardi; Hakkani-Tur, 2005), além de sistemas para recuperação de vídeo, usados para recuperação de objetos específicos (V K; GURU; Y H, 2018) e outros usos diversos (Han et al., 2017; Almeida; Leite; da S. Torres, 2011; Boureau et al., 2010).

Um desafio em comum que permeia as aplicações em recuperação de informação em diferentes modalidades trata-se de como representar um objeto multimídia. É necessário o uso de uma representação do objeto de mídia que contenha características essenciais do objeto, uma vez que trabalhar com a mídia em seu formato integral seria substancialmente custoso computacionalmente e pouco efetivo. Deste modo, a grande maioria das abordagens

utilizam de estratégias de representações vetorial de menor dimensionalidade, pois são facilmente quantificáveis e via de regra ocupam espaço de armazenamento reduzido.

As primeiras abordagens de representação de informação de imagens para recuperação se baseavam unicamente em descritores tradicionais (TORRES; FALCÃO, 2006), que possuíam como objetivo definir uma função capaz de extrair um vetor numérico a partir de uma mídia de entrada. Os descritores costumam considerar características específicas da modalidade para que foram criados, como por exemplo, descritores de imagem analisam cor, textura, contorno (PENATTI; VALLE; TORRES, 2012). Já os descritores de áudio consideram fatores como diagramas de forma de onda e espectrograma de Mel (SHARMA; UMAPATHY; KRISHNAN, 2020), enquanto descritores de vídeo consideram fatores temporais e de movimento (HU et al., 2011). Nos últimos anos uma nova categoria de descritores vem sendo amplamente utilizada, desenvolvida baseada no uso de Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNNs) (V K; GURU; Y H, 2018) tais descritores se beneficiam das diversas camadas de informação interconectadas visando tarefas de reconhecimento de padrões e aprendizado de características (DENG, 2014). A partir de um modelo treinado, tais redes extraem características relevantes para dados de interesse com contexto semelhante ao aprendido pela rede. Considerando que a rede neural foi treinada para aprender características relevantes de um cenário específico, ela possui a capacidade de quantificar a presença de tais características para novos dados, gerando para eles, vetores de características que ainda que infelizmente não sejam trivialmente interpretáveis posição a posição, sejam uma representação de características relevante para o cenário que serão empregados.

Um dos principais motivos que levaram a popularização do uso de aprendizado profundo para extração de características(*deep features*) é que abordagens baseadas em descritores clássicos possuem mais fortemente a presença de uma limitação significativa, a lacuna semântica (*"semantic gap"*) (HOI; LIU; CHANG, 2010), presente entre as características de baixo nível e conceitos semânticos de alto nível. Dado que por exemplo, ao representar uma imagem como um histograma de cores, perde-se informações de alto nível do que a imagem visava representar, tais características podem ser significativas na tarefa de recuperação de informação. A mesma limitação ocorre com a utilização de *deep features*, porém de modo menos latente, dado que as arquiteturas profundas são capazes de aprender sobre características de alto nível presentes no conjunto de dados, gerando com isso representações que serão semelhantes entre dois objetos de mesma classe semântica. Porém, ainda que presente em menor nível torna-se necessário encontrar meios de reduzir essa lacuna na recuperação, como considerando uma nova distância que leva em conta a geometria do conjunto de dados global e suas informações contextuais, ao invés de relações como distâncias par-a-par de objetos. O interesse na diminuição da lacuna semântica está no fato de que sua redução acarreta diretamente numa melhora de eficácia na recuperação de informações semelhantes, dado que serão consideradas informações com

contexto semântico mais adequadas ao conjunto de dados como um todo.

Com o objetivo de reduzir os impactos negativos da lacuna semântica nos resultados de recuperação, métodos de aprendizado não supervisionado foram propostos baseados em técnicas de *manifold learning*. Comumente métodos de aprendizado não supervisionado baseados em ranqueamento substituem distâncias entre pares de indivíduos por medidas capazes de analisar a coleção de modo global, considerando relações entre os objetos e a geometria do conjunto de dados (WANG et al., 2011) e se baseiam na ideia de recuperar as informações mais similares a partir de um objeto de entrada fornecendo uma lista ranqueada (*ranked list*). No decorrer dos últimos anos, os métodos de aprendizado não supervisionado baseados na análise das listas ranqueadas ganharam um grande destaque por possuírem baixo custo computacional e alta eficácia (PEDRONETTE; TORRES, 2017; WANG et al., 2011). Por tais métodos representarem conjuntos de dados em termos de relações de similaridade, pode-se generalizar seu uso para diversas tarefas de recuperação, inclusive as multimodais. Trabalhos anteriores já exploraram tal possibilidade, para recuperação de áudio (CAMPOS; PEDRONETTE, 2016) e vídeo (ALMEIDA; PEDRONETTE; PENATTI, 2014; ALMEIDA; VALEM; PEDRONETTE, 2017). Diferentes representações e modalidades de informação de um mesmo objeto de mídia contêm informações distintas e muitas vezes complementares. Logo, uma abordagem multimodal apresenta grande potencial, tal como já demonstrado por resultados obtidos na literatura (Yang; Meinel, 2014; Zhu; Shyu, 2015; GUO et al., 2015).

Outra abordagem que tem atraído a atenção da comunidade científica nos últimos anos são as Redes Convolucionais baseadas em Grafos (*Convolutional Neural Networks* - GCNs). As GCNs foram desenvolvidas tendo como objetivo principal lidar com problemas pertencentes ao espaço não-Euclidiano, representados na forma de estruturas de conexões não regulares entre elementos, por meio de estruturas de grafos (WU et al., 2021). Neste cenário, tais redes tem se estabelecido como uma ferramenta relevante para modelar conexões entre elementos das coleções em tarefas de recuperação, dado que ligações não diretas entre objetos de uma mesma classe semântica do conjunto podem ser aprendidas e descobertas por meio de sua aplicação. De modo simplificado, o processo de treinamento de tais arquiteturas ocorre de modo que cada nó do grafo se comunica com sua vizinhança da camada, dessa forma, cada vértice do grafo provê informação sua para a vizinhança, de mesmo modo que recebe informação da vizinhança para si, enquanto esses novos valores são atualizados para as próximas camadas da rede (ABADAL et al., 2021). Tal processo gera representações mais abrangentes e com mais informações do contexto do objeto em relação à coleção. Considerando a capacidade das GCNs em modelar a informação contextual, combinamos métodos de aprendizado não supervisionado com redes convolucionais baseadas em grafo para aprimorar a eficácia da recuperação de informação. Neste trabalho, ainda considerando a escassez de objetos de mídia rotulados disponíveis e a proibitividade do custo de rotular coleções em algumas aplicações, partimos de um

cenário não supervisionado para o treinamento das GCNs com o algoritmo *Deep Graph Infomax* (VELIčKOVIć et al., 2019).

Deste modo, considerando que informações de múltiplas modalidades provêm visões distintas e complementares dos dados que podem ser exploradas para melhorar a efetividade das representações e eficácia de tarefas de recuperação, propomos abordagens em dois cenários distintos de recuperação multimídia capazes de combinar as informações de diferentes representações. Ambas os cenários exploram estratégias de aprendizado de representações e métodos não supervisionados de *manifold learning* baseados em ranqueamento. Contudo, as abordagens diferem entre si em relação a tarefa, a forma de modelar a informação multimodal e as redes utilizadas. Enquanto o primeiro cenário compreende abordagens que utilizam de GCNs para criação de tais representações, visando a recuperação de vídeos, o segundo cenário utiliza de CNNs 3D para gerar de tais representações numa tarefa de recuperação de imagens.

O primeiro cenário de abordagens é proposto visando a recuperação de vídeos, considerando informações de diferentes modalidades (imagem, áudio e vídeo). Nesse primeiro cenário as abordagens propostas utilizam uma estratégia de *manifold learning* baseada em ranqueamento para fusão primária ou tardia (*early* ou *late*) das modalidades, além de GCNs treinadas de modo não supervisionado para a criação de representações mais abrangentes que contenham informações de vizinhança para aprimorar a recuperação. Além da avaliação de tais abordagens, apresentamos também a comparação com *baselines clássicas* para três conjuntos de dados públicos (UCF-101, MSR-VTT, TV *Human Interaction*). Em um segundo cenário, apresentamos uma abordagem de aprendizado de representações para recuperação de imagens baseado em fusão de representações multimodais, utilizando de uma CNN 3D pré-treinada para a partir dos top-k vizinhos de cada imagem gerar representações que contenham informações da imagem e de sua vizinhança. Tal abordagem também foi avaliada utilizando três conjuntos de dados públicos (Stanford 40 *Actions*, Willow *Actions* e Iklizer *Dataset*). Tendo apresentado brevemente os cenários de abordagem, é válido notar que dentre os principais diferenciais desse trabalho temos o emprego de GCNs treinadas de modo não supervisionado para recuperação de vídeo e a utilização de CNN-3Ds para a criação de representações de vizinhança para imagem.

O restante do texto está organizado da seguinte forma:

- O **Capítulo 2** apresenta a definição de aspectos teóricos relativos ao contexto do trabalho e discute trabalhos relacionados em recuperação multimodal;
- No **Capítulo 3** são apresentadas os métodos, métricas, conjuntos de dados e descritores de características utilizados em cada cenário;
- No **Capítulo 4** são discutidas as abordagens relacionadas a recuperação de vídeo, apresentando e discutindo também os resultados obtidos da avaliação de tais

abordagens.

- O **Capítulo 5** discute a abordagem de recuperação de imagens proposta, seus detalhes e resultados;
- Por fim, O **Capítulo 6** apresenta considerações finais e trabalhos futuros.

2 Referencial Teórico e Trabalhos Relacionados

Este capítulo define os principais aspectos teóricos relacionados ao tema do projeto, discutindo os conceitos essenciais de um sistema de recuperação de informações, aprendizado de representações, métodos de *manifold learning*, redes convolucionais baseadas em grafos e também apresenta uma revisão da literatura sobre os assuntos abordados. O restante do capítulo está dividido de modo que a Seção 2.1 apresenta sobre a definição formal de recuperação de informação e sistemas multimodais, enquanto a Seção 2.2 apresenta sobre diferentes abordagens de combinação empregadas na recuperação multimodal de informações, seguida da Seção 2.3 que define o termo aprendizado de representações e apresenta uma revisão de trabalhos recentes. Enquanto a Seção 2.4 apresenta uma revisão bibliográfica para a extração de características e recuperação de informações em diferentes modalidades. Por fim, a Seção 2.5 discorre sobre redes convolucionais baseadas em grafos e o algoritmo utilizado para treinamento da arquitetura utilizada.

2.1 Recuperação de Informação Multimodal

Esta seção apresenta fundamentos de Recuperação de Informação na Seção 2.1.1, enquanto discutes aspectos da recuperação multimodal na Seção 2.1.2.

2.1.1 Recuperação de Informação

Recuperação de Informação(RI) consiste em uma área significativamente abrangente da Ciência da Computação, compreendendo desde a representação, o armazenamento, a organização e o acesso a objetos de informação, tudo com o objetivo de facilitar o acesso do usuário a dados de seu interesse. Tais dados podem ser páginas Web, registros estruturados e semi-estruturados, objetos multimídia, dentre outros diversos tipos de dados. (BAEZA-YATES; RIBEIRO-NETO, 2013).

Dessa forma, um problema de recuperação de informação pode ser entendido como um abrangente espectro de aplicações, que compreendem tarefas como criar uma representação que extraia características relevantes para o conjunto de dados, calcular a similaridade entre as instâncias, efetivar etapas de pós-processamento de consultas, e diversas outras tarefas que podem compor o fluxo de execução.

Ainda nessa seção é apresentada a notação para recuperação de informações, em especial recuperação de imagens baseada no conteúdo, proposta e utilizada em trabalhos

relacionados (TORRES; FALCÃO, 2006; PEDRONETTE; GONçALVES; GUILHERME, 2018a; Carlos Guimarães Pedronette; VALEM; TORRES, 2021). As tarefas de recuperação são realizadas com base em características (*features*) extraídas dos objetos. Na atualidade, comumente são utilizados modelos profundos treinados em conjuntos de dados de grande escala para extrair características para tarefas não supervisionadas por meio de transferência de aprendizagem, isso ocorre de modo que a saída da última camada totalmente conectada do modelo pré-treinado é utilizada como um vetor de características do objeto alimentado no modelo.

Seja $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ uma coleção de objetos, onde n denota o tamanho da coleção, um extrator de características pode ser formalmente definido como uma função f. Formalmente, a função $f : \mathcal{X} \to \mathbb{R}^d$ calcula um vetor d-dimensional para um dado objeto da coleção, tal que $\mathbf{v}_{mi} = f(x_i)$ e $\mathbf{v}_m i = [v_{mi1}, v_{mi2}, \ldots, v_{mid}]$, em que d define a dimensionalidade do vetor de características.

Uma função de distância, que calcula a distância entre dois objetos de acordo com a distância entre seus vetores de características correspondentes pode ser definida como ρ: $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$. Portanto, a distância entre dois objetos x_i, x_j pode ser calculada por $\rho(\mathbf{v}_{mi}, \mathbf{v}_{mj})$. Uma tarefa geral de recuperação de informação baseada em características extraídas pode então ser modelada como o cálculo de uma lista ranqueada τ_q em resposta a um objeto de consulta x_q, de acordo com a função de distância ρ. Espera-se que as posições superiores das listas ranqueadas contenham os objetos mais relevantes em relação ao objeto da consulta. Em geral as listas são compostas de um sub-conjunto $\mathcal{X}_L$ da coleção de tamanho L, de modo que o comprimento L considerado é comumente muito menor que a colação, isto é, $L \ll n$. Também nos referimos ao conjunto de vizinhança como um pequeno conjunto de objetos similares que estejam entre os k-vizinhos mais próximos, de modo que $k \ll L \ll n$.

Formalmente, a lista ranqueada τ_q pode ser definida como uma permutação (x_1, x_2, …, x_L) do subconjunto $\mathcal{X}_L \subset \mathcal{X}$, que contém os L objetos mais semelhantes para o objeto de consulta x_q, tal que $|\mathcal{X}_L| = L$. Uma permutação τ_q é uma bijeção do conjunto $\mathcal{X}_L$ no conjunto $[n_L] = \{1, 2, \ldots, L\}$. Para uma permutação τ_q, interpretamos $\tau_q(x_i)$ como a posição (ou classificação) do objeto x_i na lista ranqueada τ_q. Se x_i é posicionado antes de x_j na lista classificada de x_q (ou seja, se $\tau_q(x_i) < \tau_q(x_j)$), então ρ ($\mathbf{v}_q$, $\mathbf{v}_{mi}$) $\leq$ ρ ($\mathbf{v}_q$, $\mathbf{v}_{mj}$).

Tomando cada objeto $x_i \in \mathcal{X}$ como um objeto de consulta x_q, um conjunto de listas ranqueadas $\mathcal{T}$ pode ser calculado, contendo uma lista ranqueada para cada objeto na coleção. Cada lista ranqueada estabelece uma relação de similaridade entre o objeto de consulta e todos os objetos na coleção $\mathcal{X}$. Portanto, o conjunto $\mathcal{T}_m$ codifica uma rica fonte de informações de similaridade/dissimilaridade sobre a coleção $\mathcal{X}$.

2.1.2 Sistemas de Recuperação Multimodal

Esta seção tem como objetivo definir um modelo de recuperação multimodal genérico baseado em ranqueamentos, discutindo as principais etapas comumente presentes em tais sistemas. A Figura 1 ilustra as principais etapas em uma arquitetura genérica de modelo de recuperação multimodal. A representação gráfica das etapas está identificada de acordo com a numeração referenciada no texto a seguir.

De modo geral, o processo inicia pela extração de características de toda a coleção de dados para cada modalidade (Passo **I**), que ocorre de modo análogo ao modelo de apenas uma modalidade, em que seja $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ uma coleção de objetos, onde n denota o tamanho do coleção, m diferentes descritores tais que contemplam funções do modelo f_m são empregados, de modo que geram m vetores de características para cada objeto, tais que $\mathbf{v}_{mi} = [v_{mi1}, v_{mi2}, \ldots, v_{mid_m}]$ em que d_m refere-se ao tamanho do vetor de características para a modalidade m.

Uma vez realizada a extração de características ocorre o cálculo da distância entre as características de cada objeto para cada modalidade (Passo **II**), de modo que sendo ρ_m: $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$, a função de distância entre dois objetos x_i, x_j pode ser calculada por $\rho_m(\mathbf{v}_{mi}, \mathbf{v}_{mj})$ para a modalidade m. Gerando com isso m conjuntos de listas ranqueadas $(\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_m)$ (Passo **III**) tal que em cada conjunto cada objeto q possui uma lista ranqueada τ_{mq} na modalidade m.

Por fim, considerando que as múltiplas modalidades ocasionam em múltiplos conjuntos de listas ranqueadas torna-se necessário um método de combinar as informações desses múltiplos ranqueamentos (Passo **(IV)**, com o objetivo de produzir um ranqueamento final que possua melhor eficácia de recuperação em relação aos outros. Tal processo pode ser definida como uma função genérica g_S detalhada na Equação a seguir:

$$\mathcal{T}_n^* = g_S(\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_m) \tag{2.1}$$

Onde $\mathcal{T}_n^*$ denota o conjunto de ranqueamentos selecionados pela abordagem, para um determinado tamanho n, tal que $|\mathcal{T}_n^*| = n$.

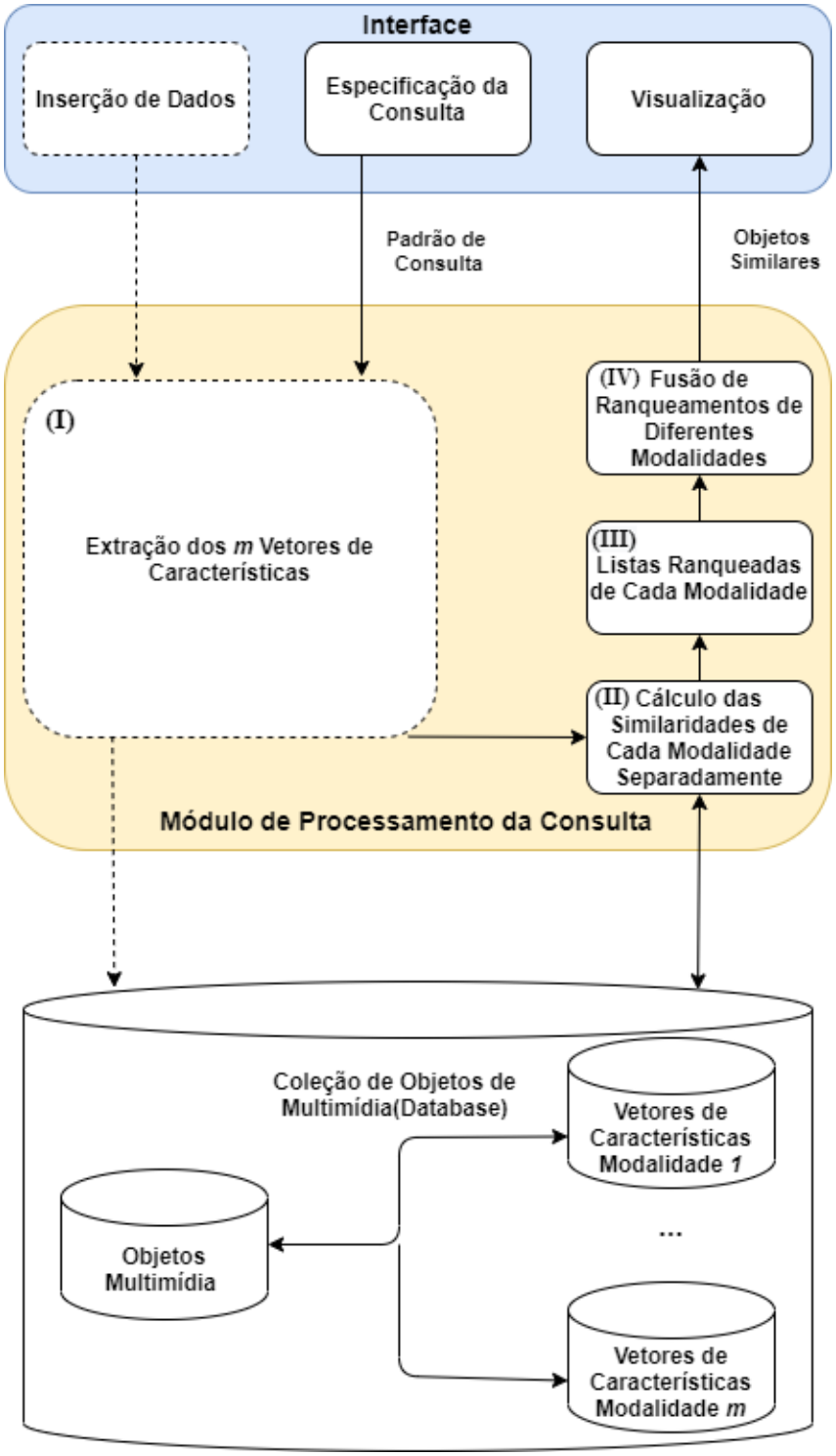


Figura 1 – Arquitetura genérica de um modelo de recuperação multimodal.

2.2 Estratégias de Combinação

Diversas técnicas para combinação de informações de diferentes descritores têm sido propostas na literatura (BHOWMIK et al., 2014), via de regra adaptadas especificamente para os descritores selecionados e o cenário específico de utilização. De modo geral, as estratégias podem ser organizadas em três categorias principais: fusão precoce, fusão tardia

e fusão intermediária (SNOEK; WORRING; SMEULDERS, 2005; ATREY et al., 2010). Uma estratégia de combinação é caracterizada entre tardia ou precoce de acordo com em qual etapa ela ocorre, como pode ser visto na Figura 2.

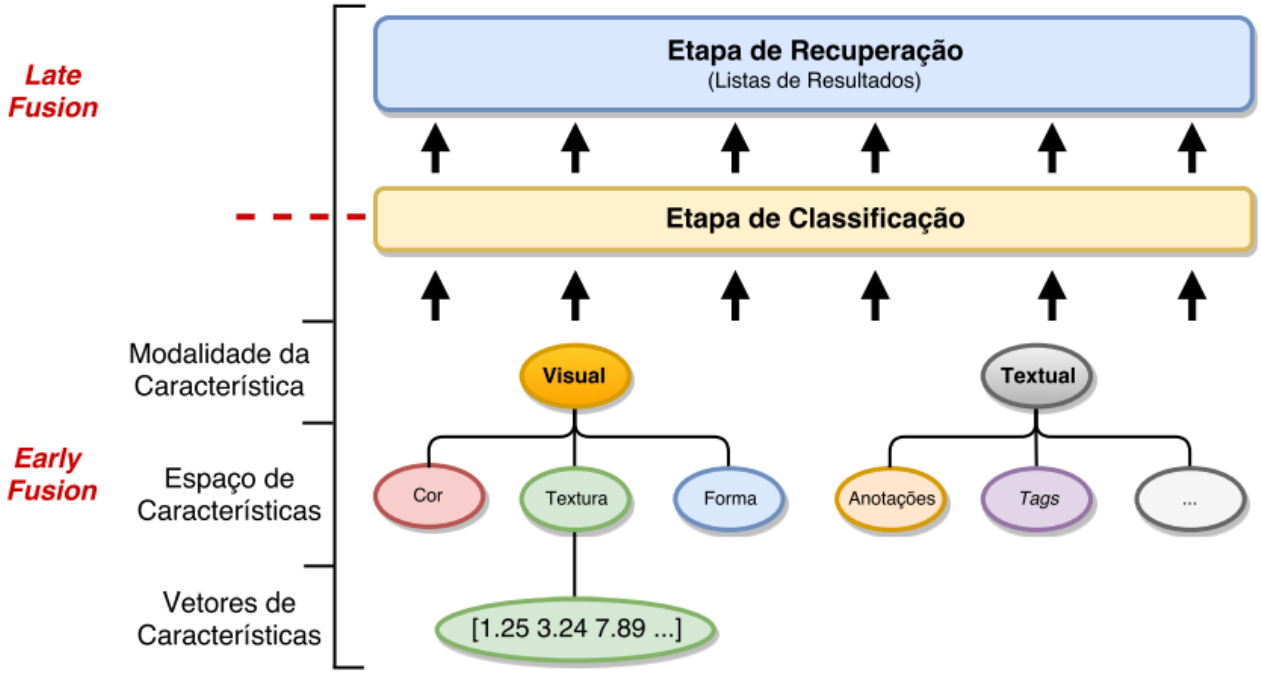


Figura 2 – Representação de etapas onde a fusão de características representa um fusão tardia ou uma fusão precoce. Adaptada de (PIRAS; GIACINTO, 2017)

2.2.1 Fusão Tardia

Nas abordagens baseadas em fusão tardia, a combinação de diferentes características ocorre depois do cálculo da distância dos objetos da coleção. Deste modo, a combinação não é realizada sobre vetores de características diretamente, dado que os mesmo já foram processados em alguma representação de similaridade. Métodos de re-ranqueamento (VALEM; PEDRONETTE, 2016), como os baseados em vizinhança e aprendizado não supervisionado, comumente realizam formas de fusão tardia. Outra abordagem trata-se de atribuir pesos distintos a cada modalidade de dados a serem combinados (ZHANG; QIN; WAN, 2011; PIRAS; TRONCI; GIACINTO, 2013). A Figura 3 apresenta um esquema de uma sequência de passos para uma recuperação multimodal genérica que utiliza de fusão tardia.

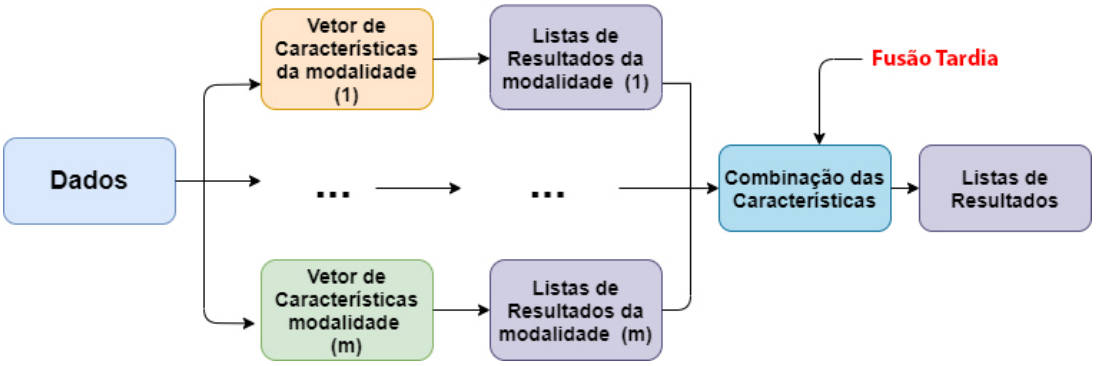


Figura 3 – Modelo de combinação baseado em fusão tardia.

2.2.2 Fusão Precoce

Nas abordagens baseadas em fusão precoce, ocorre a combinação de diferentes características em uma única representação, em etapa anterior ao cálculo de distância entre os elementos do conjunto de dados, ou criação de ranqueamentos (SNOEK; WORRING; SMEULDERS, 2005). Tal modalidade assim como a fusão tardia, tende a ser comum em aplicações de recuperação de conteúdo, comumente a fusão precoce ocorre com a concatenação de vetores de características de múltiplas modalidades de um objeto em um único vetor (YU et al., 2013). (YUE et al., 2011) realizam a fusão precoce de descritores baseados em diferentes representações de imagens (unindo as características de cor, forma e textura). Na Figura 4 é apresentada um fluxo de passos genéricos de um sistema de recuperação de informações multimodal onde ocorre fusão precoce.

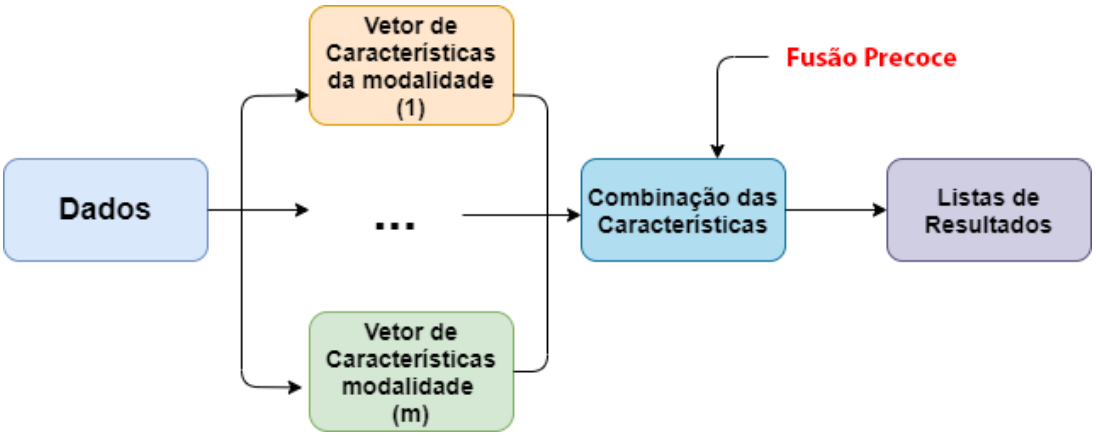


Figura 4 – Exemplo de modelo com presença de fusão precoce.

2.2.3 Fusão Intermediária

Uma terceira categoria não tão amplamente utilizada trata-se da fusão intermediária (IONESCU et al., 2014), a qual é diferente da fusão híbrida, em que utiliza da combinação da fusão tardia e precoce. Na fusão intermediária, vetores de características extraídos do conjunto de dados são utilizados como entrada para treinamento de classifica-

dores, como SVM(*Support Vector Machine*) ou redes profundas, os classificadores são então treinados para escolher a melhor combinação entre os vetores de entrada, criando uma nova representação que combina múltiplas modalidades. Como o treinamento ocorre sobre os vetores de características visando encontrar uma nova representação, essas abordagens às vezes na literatura são referenciadas como fusão precoce, porém por possuírem a presença de um classificador de aprendizado profundo em meio a fusão, via de regra essas abordagens são atribuídas a essa terceira nova categoria de estratégias de combinação.

2.3 Aprendizado de Representações

Aprendizado de representações trata-se de uma área consideravelmente abrangente, dado que compreende todos os métodos que utilizam de aprendizado de máquina para gerar novas representações para dados de entrada. Dessa forma, o conceito pode compreender desde abordagens de aprendizado supervisionado baseadas em modelos de aprendizado profundo utilizadas para gerar representações, como também métodos de aprendizado não supervisionado, *manifold learning* e redução de dimensionalidade que representam os objetos da coleção em um novo espaço (BENGIO; COURVILLE; VINCENT, 2013).

Uma abordagem de aprendizado de representações trata-se de usar coleções de objetos como dados de entradas em modelos pré-treinados de aprendizado profundo, a fim de se obter uma nova representação da informação de entrada. Esse procedimento é feito considerando que a nova representação evidenciará conceitos relevantes de alto nível aprendidos pelos modelos pré-treinados e até então não quantificados ou não processados nos objetos originais. Recentemente esse conceito vem sendo bastante aplicado, como realizado por Wiggers et al. (2019), que a partir de uma CNN pré-treinada no conjunto de dados ImageNet (DENG et al., 2009), realizam a recuperação de documentos antigos e também a identificação de padrões a partir das representações geradas pelos modelo profundo, tal abordagem obteve resultados significativamente positivos de Precisão e MAP em recuperação. De modo semelhante, Rian, Christanti e Hendryli (2019) utilizam de uma rede neural convolucional pré-treinada no ImageNet, para extração de representações, a qual em conjunto de uma SVM realiza a recuperação de imagens para o conjunto de dados ImageNet.

Ainda nessa modalidade de aprendizado de representações, Chung e Weng (2017) propõem um modelo para aprendizado de representações com a utilização de ResNets(Residual Networks) focado na recuperação de imagens médicas para identificação de retinopatia diabética. Em seu método utilizam de uma CNN siamesa, de modo que duas ResNets possuem suas saídas conectadas a um bloco que calcula a distância entre as saídas das duas redes, que são alimentadas com instâncias diferentes ao mesmo tempo, visando aprender uma representação do conjunto de dados em termos de distância entre as instâncias que o

compõe.

Em uma abordagem de aprendizado de representação multimodal Ning, Zhao e Yuan (2021) realizam o uso de uma VGG e de uma CNN pré-treinada para extrair de dados de sensoriamento remoto, sendo respectivamente representações visuais e de áudio, de modo que realizam a combinação de ambas representações a fim de obterem uma representação semanticamente mais consistente e de melhor eficácia para recuperação.

Varamesh et al. (2020) apresentam um *framework* para aprendizado de representações que utiliza de aprendizado auto-supervisionado e se baseia em ranqueamentos para recuperação de imagens. De modo que um codificador de representações é treinado visando maximizar a precisão média(*Average Precision*) do ranqueamento em uma abordagem nomeada S2R2, visando alcançar o conjunto de ranqueamentos ideais para a coleção de dados.

A abordagem de aprendizado de representações para recuperação de imagens desse trabalho utiliza de modelos pré-treinados para recuperação de imagens, em que combinamos informações de ranqueamento obtidas a partir de uma CNN 2D com informações obtidas de uma representação extraída de uma CNN 3D. Pode-se destacar como diferencial em relação aos trabalhos relacionados o fato de que a representação do modelo 3D contém não apenas informações da imagem de consulta, mas também de sua vizinhança.

2.4 Recuperação Baseada em Diferentes Modalidades

Esse seção discute diversas abordagens de recuperação de conteúdo, considerando diferentes modalidades e também cenários de multimodalidade. O Capítulo está divido de modo que a Seção 2.4.1 apresenta um levantamento bibliográfico quanto à extração de características e recuperação de imagens, enquanto a Seção 2.4.2 apresenta discussão análoga para vídeos. Por sua vez, a Seção 2.4.3 faz o equivalente para a modalidade de áudio.Por fim a Seção 2.4.4 apresenta diferentes abordagens propostas para sistemas de recuperação multimodal.

2.4.1 Recuperação de Imagens

Uma representação que considera apenas os *pixels* de uma imagem sem qualquer pré-processamento, em geral gera resultados pouco eficazes em tarefas de recuperação e reconhecimento de padrões. Dessa forma, os descritores visuais são propostos com a finalidade de extrair características relevantes e armazená-las em vetores de características (DATTA et al., 2008). Dentre as abordagens de descritores para imagens temos três principais categorias, descritores locais, descritores globais e um nicho de abordagem significativamente mais recente baseado em aprendizado profundo. A divisão entre as duas primeiras categorias vem de como eles analisam a imagem, pois enquanto os globais

descrevem toda a imagem, os locais descrevem apenas partes representativas da imagem, por meio das quais realizam o processo de *matching* entre imagens. Já a terceira categoria engloba abordagens que utilizam de modelos profundos para extração de características.

Dentre os descritores globais para imagens temos três principais categorias que são as baseadas em cores, as baseadas em texturas e as baseadas em formas. A extração de características baseadas em cores depende do espaço de cores utilizado (RGB, HSV, YCbCr, YUV, HVC), que por sua vez depende da aplicação utilizada. Características de cores englobam representações como histogramas de cores, que visam identificar cores mais e menos frequentes, correlação entre cores que visam encontrar relações entre a presença de determinadas cores e modelos Gaussiano. Como alguns exemplos de descritores baseados em cores na literatura temos o *Color Autocorrelogram* (HUANG et al., 1997) e o *Global Color Histogram* (SWAIN; BALLARD, 1991). Quanto aos pontos negativos de abordagens baseadas apenas em descritores de cores está que elas não consideram textura, forma e informações de alto nível, ignorando com isso informações significativamente importantes.

Descritores baseados em textura de modo geral tratam de características intrínsecas da superfícies de objetos presentes na imagem, contendo informações cruciais sobre a organização das superfícies e suas correlações com o ambiente. De modo geral ocorre a construção de modelos, treinados para identificar diferentes tipos de superfícies (como quadriculados, superfícies porosas, asfalto,etc.) e são utilizadas como entrada nesses modelos as imagens que se objetiva extrair os vetores de características. Como exemplos desse tipo de descritor temos o *Local Binary Patterns* (OJALA; PIETIKäINEN; MäENPää, 2002), o *Color Co-Ocorrence Matrix* (KOVALEV; VOLMER, 1998) e o *Local Activity Spectrum* (TAO; DICKINSON, 2000). Dentre os benefícios de características de textura, está que podem ser aplicadas em qualquer aplicação onde a textura da imagem seja relevante e presente, mas pelo lado negativo nem todos as imagens possuem resolução suficiente para que a textura seja avaliada de modo eficaz, além de se tratar de descritores que também não consideram características de alto nível.

Descritores de forma têm como objetivo descrever a forma dos objetos que compõem a imagem, podendo ser extraída de contornos de objetos ou das regiões que esses contornos delimitam. Uma abordagem comum é detectar as aresta s(comumente utilizando algoritmos como o *Canny edge detection*) presentes na imagem. Dentre os exemplos desse tipo de descritores na literatura temos o *Aspect Shape Context* (LING; YANG; LATECKI, 2010), o *Segment Saliences* (TORRES; FALCÃO, 2007) e o *Countour Features Descriptor* (PEDRONETTE; TORRES, 2010). Abordagens baseadas em forma são mais efetivas em aplicações em que a informação da forma está explicita e legível, algo que pode não ocorrer em imagens com presença de borrões, ruídos ou de baixa resolução.

Quanto a descritores locais, uma abordagem que se destaca e tem sido bastante utilizada na literatura trata-se da BoVW(*bag of visual words*), inspirada no BoW(*bag*

of words), técnica utilizada na recuperação de textos. Para imagens, essa abordagem se baseia em analisar diferentes imagens procurando pontos chave de alta relevância para descrever as imagens e que ainda assim sejam invariantes, esse pontos nada mais são do que pedaços das imagens que podem possuir tamanhos distintos, porém visam descrever algo como um objeto específico (LOWE, 1999). Com essas partições da imagem é construído um dicionário de palavras visuais. E então um vetor de características pode ser construído considerando a ocorrência de cada um dos elementos do dicionário em uma abordagem com reconhecimento de objetos por exemplo. Dentre os principais descritores dessa modalidade temos o *Scale-Invariant Feature Transform*(SIFT) (LOWE, 1999), *Speeded Up Robust Features*(SURF) (BAY et al., 2008), o *Binary Robust Independent Elementary Features*(BRIEF) (CALONDER et al., 2010) e o *Oriented Fast and Rotated* BRIEF(ORB) (RUBLEE et al., 2011). Uma abordagem que utiliza de descritores locais (mais especificamente o SIFT) pode ser visualizada na Figura 5, onde temos dois objetos em posições e escalas diferentes, porém a combinação(*"matching"*) dos pontos permite a identificação do objeto. As abordagens de BoVW foram amplamente utilizadas no campo de recuperação, de modo a ser predominante até desenvolvimento e popularização das abordagens baseadas em aprendizado profundo (SALAU; JAIN, 2019).

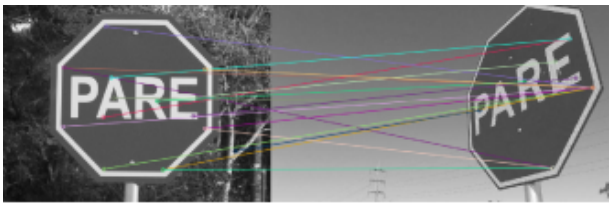


Figura 5 – Exemplo de uso de descritores locais com dicionário de palavras.

Recentemente tem sido explorada uma nova abordagem para extração de características de objetos multimídia em geral, baseada em modelos de aprendizado profundo (HAYAKAWA et al., 2016), (SURESHA; KUPPA; RAGHUKUMAR, 2020). Em que comumente são empregadas CNNs(*Convolutional Neural Networks*), dado o notável crescimento da pesquisa sobre esse tipo de arquitetura nos últimos anos (SERMANET; CHINTALA; LECUN, 2012) e robustos resultados alcançados. De modo geral, essa abordagem parte da introdução da instância, ou de uma representação da instância, como entrada em um modelo profundo. Esse modelo, por sua vez, trata-se de uma arquitetura pré-treinada de domínio relevante ao que a imagem de entrada pertence e por meio dele é extraído um vetor de características, comumente da camada anterior a que gera as probabilidades de classificação ou regressão do modelo. Dentre as causas de sucesso deste modelo, pode-se destacar a capacidade de camadas anteriores à classificação em extrair características genéricas relevantes. Tal processo dá origem a um vetor que ainda que não seja trivial de ser analisado contextualmente posição a posição, seja uma representação numérica simples de ser trabalhada e semanticamente representativa de tal modo que grande parte

dos problemas relacionados a entendimento de vídeo, imagem e áudio atualmente utilizam de redes neurais profundas para resolução (LI; DENG, 2018).

Apesar do desenvolvimento de descritores baseados em aprendizado profundo os descritores clássicos (globais e locais) ainda apresentam cenários de usdo na literatura para recuperação de informações. Como demonstrado no trabalho de Baji e Mocanu (2018), que realizam a recuperação de imagens utilizando de componentes conectados e características de forma, gerando um histograma e vetores de características estatísticos. As características de cor e textura dos componentes conectados são então utilizados para o cálculo da matriz de co-ocorrência de tons de cinza que são utilizados para recuperar imagens de um grande banco de dados de imagens. Outra abordagem também utilizando de características extraídas a partir de descritores clássicos, nesse caso locais, trata-se do trabalho realizado por Benam, Drew e Atkins (2017) que propõem um sistema de recuperação de imagens dermatoscópicas (imagens de pele capturadas com um dermatoscópio) baseando-se em uma rede de pigmentação. O sistema recebe uma imagem de entrada e retorna as imagens com redes de pigmentação mais semelhantes a imagem de consulta. Nessa abordagem para cada imagem, pontos de interesse para a Dermatoscopia são detectados e um vetor de de 128 características é extraído. Então os vetores são recuperados de acordo com um algoritmo simples de compatibilidade com pesos.

Além dos descritores, os sistemas de recuperação de imagem também evoluíram ao longo dos últimos anos em outros aspectos e direções, não apenas ligadas ao desenvolvimento de descritores. Quanto a trabalhos de recuperação de imagens baseados em modelos profundos, temos abordagens como a de Babenko et al. (2014) que utilizam as ativações das camadas do topo de uma profunda rede neural convolucional como descritores de características para recuperação de imagens. Tal estudo apresentou uma performance significativamente promissora para um modelo treinado com dados do mesmo domínio da coleção de dados. Outro exemplo é o trabalho realizado por Wang et al. (2014) que utiliza de um modelo de ranqueamento profundo para a recuperação, criando uma métrica de similaridade que é diretamente aprendida a partir das imagens de entrada. Deste modo um algoritmo de amostragem tripleto(composto de instâncias que são tuplas com a imagem de busca, a imagem mais semelhante da mesma classe e a imagem menos semelhante da mesma classe) é proposto para o modelo aprender ranqueamentos para melhorar a recuperação. Forcén et al. (2020) utilizam da saída da ultima camada de convolução de uma CNN como uma representação para modelar co-ocorrências a partir das *deep features*. A partir dessa modelagem de co-ocorrências e mapas de características é criada uma representação de imagens, utilizada para a recuperação de imagens.

Dentre abordagens utilizando aprendizado não supervisionado, temos trabalhos como o de VALEM L. P. (2016), que a partir da aplicação do método de aprendizado não supervisionado Produto Cartesiano de Referencias de Ranqueamento, explora as múltiplas

informações intrínsecas nos ranqueamentos de similaridade dos vetores de características, aumentando significativamente a eficácia de recuperação. Também utilizando de aprendizado não supervisionado, Pedronette, Gonçalves e Guilherme (2018b) aplicam aprendizado não supervisionado para através de um grafo kNN(*k-Nearest Neighbors*) e componentes conectados realizarem tarefas de recuperação de imagens, de modo que o método explora a geometria intrínseca do conjunto de dados para aprimorar a eficácia da recuperação de imagens, gerando um novo ranqueamento que não apenas considera distâncias par a a par mas toda a coleção, representando uma estratégia de baixo custo computacional e altos ganhos de eficácia em comparação a métricas de distância tradicionais.

2.4.2 Recuperação de Vídeos

Para a extração de vídeos temos também a divisão entre descritores locais e globais e dentro de globais temos as subcategorias de características estáticas e características baseadas em movimento. As características estáticas se assemelham as empregadas para imagens, visto que analisam informações de quadros chave estaticamente.

Dentre os descritores locais, de modo semelhante a imagens, a abordagem que mais se popularizou foram as de estratégias baseadas em *Bag of Visual Words* (BoVW) (Boureau et al., 2010), que trata-se de realizar a representação do video utilizando da informação estatística de padrões locais, codificando as ocorrências de características locais de modo quantizado. Essa codificação cria o que é chamado de dicionário visual ou livro de código visual, após isso ocorre a codificação e pooling (Boureau et al., 2010) do mesmo. Outra abordagem local relacionada, trata-se da *Bag of Scenes* (Boureau et al., 2010), um dicionário composto das cenas de interesse, porém contendo apenas uma representação de características locais.

Os descritores de características estáticas são dividas entre baseadas em cor, textura e forma. Características baseadas em cores incluem histograma de cores, momentos de cores, correlação entre cores, dentre outros. Dentre alguns trabalhos que utilizam esses descritores temos Amir et al. (2003) que utilizam do histograma de cores e momentos de cores para realizar recuperação de vídeo e detecção de conceitos. Também utilizando de características de cores temos Yan e Hauptmann (2007) que primeiro dividem os quadros chave em blocos de 5×5 e após isso realizam a extração de característica de cores, realizando para cada bloco o processo de histograma de cor e momentos de cor e posteriormente utilizam para classificação. Adcock et al. (2004) utilizam diagramas de correlação de cor para implementar uma biblioteca de busca de vídeos. Dentre os méritos das características baseadas em cor está que elas refletem a percepção visual humana, são facilmente extraídas e de pouca complexidade computacional. Quanto ao pontos negativos, de mesmo modo que para imagens abordagens baseadas apenas em cores não descrevem textura e forma, desconsiderando informações significativamente importantes.

Ainda sobre características estáticas temos as baseadas em textura, análogas as de imagem. Características de textura incluem modelos simultâneos auto-regressivos, características de orientação, características de textura baseadas em transformação *Wavelet*, matrizes de coocorrência, e outros. Amir et al. (2003) utilizam coocorrência de textura, incluindo aspereza, contraste e direcionalidade para a tarefa TRECVid-2003 de recuperação de vídeo, que se refere a um desafio de recuperação sobre um conjunto de dados com cerca de 62 horas de vídeos, contendo 133 conceitos diversos a serem utilizados para recuperação. Hauptmann et al. (2003) utilizam filtros Gabor Wavelet para capturar informações de textura em busca de vídeos.

Também há abordagens baseadas em propriedades de forma. Em geral, essas abordagens consideram contornos de objetos ou das regiões que esses contornos delimitam. Hauptmann et al. (2003) utilizam de um descritor de histograma de arestas para capturar a distribuição espacial das bordas em tarefas de busca de vídeo. Em abordagens semelhantes Cooke et al. (2004) e Foley et al. (2005), primeiro dividem quadros chaves em blocos e então extraem os histogramas de aresta de cada bloco. Abordagens baseadas em forma são mais efetivas em aplicações em que a informação da forma está explícita e legível, algo que pode não ocorrer em vídeos com alto grau de movimentação, ou presença de borrões de movimento.

Por fim, dentre as abordagens tradicionais, torna-se essencial considerar também características de movimentação. Por se tratar da modalidade de vídeo dado, a dimensão temporal contém significativas informações, que estão comumente associadas à semântica do mesmo. A movimentação presente no vídeo inclui a movimentação do plano de fundo, causado pela movimentação da câmera e a movimentação do plano principal causada pela movimentação de objetos. Deste modo características baseadas em movimentação podem ser divididas entre relacionadas a câmera e relacionadas a objeto (KAUR; KAUR, 2015). Para as relacionadas a câmeras temos operações como *zoom in, zoom out*, inclinação e rotação. Porém, a utilização de apenas características de movimentação da câmera possui significativas limitações quanto a representação do conteúdo presente no vídeo e por tal razão é pouco utilizada sozinha. Quanto a área de características baseadas em movimentação de objetos, ela pode ser dividia em três categorias principais: baseada em estatística, baseada em trajetória e baseada na relação espacial dos objetos presentes no vídeo.

Características baseadas em estatísticas compreendem os movimentos de pontos específicos ao decorrer dos quadros de um vídeo, as quais são extraídas para modelar a distribuição global ou local de movimentação do vídeo. Fablet, Bouthemy e Perez (2002) utilizam de um modelo estocástico, para representar a distribuição espaço-temporal de medidas de movimentação local, calculadas estimando as principais movimentações presentes na sequencia de quadros do vídeo. Tal representação foi posteriormente empregada

na recuperação e indexação de vídeos.

Também utilizando de características estatísticas, Ma e Zhang (Yu-Fei Ma; Hong-Jiang Zhang, 2002) transformaram o campo de vetor de movimento em um número de porções direcionais de acordo com a intensidade do movimento, essas porções representam um conjunto de momentos que formam um vetor tridimensional chamado de vetor de textura de movimento, utilizado mais tarde na recuperação de instâncias com movimentação semelhante. Dentre as vantagens das características baseadas em estatísticas estão o baixo custo computacional, porém dentre as limitações temos que elas não podem representar as ações relacionadas aos objetos presentes no vídeo de modo preciso e tão pouco caracterizar relações entre objetos.

Características baseadas em trajetória são comumente extraídas modelando o movimento das trajetórias de cada objeto no vídeo. Bashir, Khokhar e Schonfeld (2007) apresentam um mecanismo de recuperação de vídeos baseada na movimentação da trajetória dos objetos do vídeo, em que as trajetórias são representadas por sub-trajetórias temporalmente organizadas. Jung, Lee e Ho (2001) baseiam seu modelo de trajetória no polinômio de ajuste da curva, e utilizam tal modelo como uma chave de indexamento para acessar os objetos. Su et al. (2007) constroem fluxos de movimentação a partir dos vetores de movimento para gerar informações de movimento contínuo na forma de uma trajetória, de modo que dado uma trajetória o sistema recupera o conjunto de trajetórias mais similares a ela. Dentre as limitações de características baseadas em trajetória temos que elas dependem da correta segmentação de objetos, rastreamento de movimento e representação de trajetória, três tarefas consideravelmente complexas.

Características baseadas na relação entre objetos descrevem a relação espacial entre eles. Yajima, Nakanishi e Tanaka (2002) consulta os movimentos de múltiplos objetos e especifica a relação espaço-temporal dentre eles expressando os traços de cada objeto em uma linha do tempo. Dentre os méritos de características baseadas na relação entre objetos está que elas podem de modo intuitivo representar as relações de múltiplos objetos no domínio temporal, porém dentre as limitações está a dificuldade de rotular cada objeto e posição.

Recentemente, de forma análoga ao que ocorre para imagens e áudios, novas abordagens têm sido exploradas para extração de características baseada em modelos de aprendizado profundo (HAYAKAWA et al., 2016), (SURESHA; KUPPA; RAGHUKUMAR, 2020). Nessa modalidade, comumente são empregados modelos com artifícios espaço-temporais, para que considerem também informações da dimensão de tempo e gerem características espaço-temporais. Os motivos do sucesso de tal abordagem são análogos à aplicação realizada em imagens, assim como as etapas principais envolvidas no processo.

Diversas abordagens baseadas em Redes Neurais Profundas para entendimento e extração de características de vídeos foram propostas ao longo dos últimos anos. Como por

exemplo utilizando de FlowNets (um modelo baseado em CNNs para explorar características temporais de movimento), como proposto por Hui, Tang e Loy (2018) e Ilg et al. (2017). Além de modelos Two-Stream (SIMONYAN; ZISSERMAN, 2014; ZHANG et al., 2019) em que ocorre uma *pipeline* de dois modelos de CNNs utilizadas para identificar padrões espaço-temporais a partir de uma sequência de quadros. Ou utilizando de modelos híbridos baseados em implementações Two-Stream, porém com a utilização de RNNs(*Recurrent Neural Networks*) 3D e camadas LSTM(*Long Short-Term Memory*), visando analisar informações de movimento de longo termo em relação a duração do vídeo((GAMMULLE et al., 2017; HOU; CHEN; SHAH, 2017; PENG; ZHAO; ZHANG, 2017; Singh et al., 2016)).

2.4.3 Recuperação de Áudio

A extração de características para áudio se baseia em uma abordagem significativamente diferente em relação a imagens e vídeos, pois não possui informações visuais. Suas características se baseiam em outros domínios, comumente pertencentes a quatro categorias: domínio temporal, domínio de frequência, domínio de tempo-frequência e características profundas.

Dentre os domínios o mais antigo e simples trata-se do domínio temporal, utilizado desde a década de 1950 (SMITH, 1951; GOLDMAN-EISLER, 1958; STEVENS, 1950) e representa um importante leque de características para análise de áudio e classificação. Para analisar o espectro de sinais de áudio, diversas representações são consideradas tais como as baseadas em amplitude (MITROVIC; ZEPPELZAUER; BREITENEDER, 2010), baseadas em energia, dentre outras. Características baseadas em amplitude são via de regra análises da representação temporal do sinal de áudio, comumente utilizadas em discriminação e classificação de sons de ambiente (MITROVIC; ZEPPELZAUER; BREITENEDER, 2006). Características baseadas em energia são características como volume e centroides temporais, comumente utilizadas para detectar segmentos de presença e ausência de voz (Yang et al., 2010), sons de ambiente (PELTONEN et al., 2002) e também discriminar voz e música (Fu et al., 2011).

Graças às limitações contidas nas abordagens baseadas em tempo, a partir de aproximadamente 1960, passou-se popularizar a análise de características dos sinais em termos de frequência, que pode ser obtida utilizando da transformada de Fourier ou análise da auto-regressão (predição linear de um sinal) sobre o sinal de domínio de tempo (HOWARD, 1956; STEVENS, 1950). Dentre as características de frequência, temos as baseadas em auto-regressão como Codificação Preditiva Linear, que tenta remover redundâncias do sinal e predizer os próximos valores por combinação linear, CELP (*Code Excited Linear Prediction*) que possui o mesmo principio de predição linear porém utilizando de um dicionário de entradas. Comumente tal tipo de características é empregado para

reconhecimento e classificação de sons ambientes (Tsau; Kim; Kuo, 2011).

Quanto ao domínio tempo-frequência, ele surgiu visando unir informações de ambos os domínios, visto que o domínio do tempo mostra variações do sinal quanto a amplitude em relação ao tempo, porém não possui quaisquer informações sobre frequência, e o domínio de frequência, por sua vez, possui a magnitude do conteúdo da frequência, porém nenhuma informação direta em relação ao tempo. Características baseadas na união de informações de tempo e frequência, são via de regra uma transformação do sinal de modo que podemos analisar o sinal em relação ao tempo em um eixo e em relação de frequência no outro. Comumente chamadas de representação de tempo-frequência, o modo mais comum de se obter uma representação de tempo-frequência trata-se de aplicar um procedimento baseado na *Short-time Fourier Transform*(STFT) (SEJDIC; DJUROVIC; JIANG, 2009), uma transformada utilizada para determinar a frequência senoidal e o conteúdo de fase de seções locais de um sinal de áudio ao longo do tempo. Partindo da aplicação de STFTs, obtemos características como matrizes de tempo-frequência(MTF) e Envelope de Espectro. Matrizes de tempo frequência são representações obtidas utilizando de STFTs sobre a informação de domínio temporal, porém são uma representação que gera uma quantidade de dados exponencia. Características dessa categoria possuem uma ampla gama de aplicações tal como análise de sinais de áudio, classificação de sons de ambiente (SHARMA; UMAPATHY; KRISHNAN, 2020), reconhecimento de gênero de música (LI; OGIHARA; LI, 2003) e reconhecimento de sons ambiente (PELTONEN et al., 2002).

Há também as características de domínio Cepstral, onde as características funcionam sobre a representação de cepstrum, que nada mais é do que o resultado obtido tomando a transformada inversa de Fourier do logaritmo no espectro do sinal. Dentre elas, a mais popularmente utilizada trata-se da representação Coeficiente Cepstral de Frequência de Mel (MFCCs), derivada da representação cepstral de um clipe e que visa representar o espectro de potência de um áudio em pequenos intervalos de tempo, com base na transformada discreta de cosseno do espectro de potência logarítmica em uma escala de Mel não linear. Em uma MFCCs as bandas de frequência são igualmente espaçadas na escala de mel, imitando a audição humana e por tal sendo amplamente utilizada em abordagens de reconhecimento e aprimoramento de fala (Davis; Mermelstein, 1980; Krueger; Haeb-Umbach, 2010).

Assim como para imagem e vídeo, abordagens baseadas em aprendizado profundo para extração de características têm sido significativamente presentes na pesquisa para extração de características de áudio (SHARMA; UMAPATHY; KRISHNAN, 2020). De forma análoga ao domínio de imagens, características de alto nível são extraídas de informações de baixo nível, obtidas através das saídas das camadas ocultas de modelos de redes neurais profundas em que o sinal de áudio ou uma representação dele é alimentado

em um modelo pré-treinado. Li Yanxiong et al. (2018) utilizam de características extraídas de modelos profundos para detectar eventos de som, enquanto Zhang et al. (2018) utilizam delas para classificação e recuperação de cenas a partir do áudio.

Dentre as abordagens relevantes de recuperação por conteúdo para áudio temos trabalhos como o de Foote (1999), em que a partir de uma representação própria de objetos de áudio que se baseia nas características MFCC e de energia, são gerados vetores de característica. Sobre esse vetores, com a utilização de medidas de distância tradicionais (Euclidiana e Cosseno) são geradas listas ranqueadas. Outra abordagem também baseada em MFCCs é realizada por Vaidya e Shah (2014), porém dessa vez com uma representação que une características de *Pitch*,*loudness* e Tom. Pitch trata-se do quão alto ou baixo o som é, enquanto *loudeness* por sua vez é a percepção subjetiva de pressão sonora e Tom se refere à altura de um som na escala geral de sons.

Dentre as abordagens utilizando métodos de aprendizado não supervisionado temos trabalhos como o de (Panyapanuwat; Kamonsantiroj; Pipanmaekaporn, 2019), que utilizam de modelos profundos para aprender uma representação de *hashing* não supervisionada a partir dos áudios de entrada e que a partir da função de *hashing* recupera objetos semelhantes ao de consulta. Xue-Yan Zhao, Fei Wu e Jie Lin (2004) realizam em seu trabalho o emprego de métodos de aprendizado não supervisionado para recuperação de áudio, mais especificamente do método de agrupamento *Minimum Spannig Tree*(MST) para realizar o *matching* de áudios, em que para realização da recuperação são aplicados pesos diferentes para cada característica, aprendidos de modo não supervisionado a partir do valor dos centroides dos agrupamentos de instâncias.

2.4.4 Recuperação Multimodal

Abordagens que utilizam de múltiplas modalidades de informação para recuperação são nomeadas abordagens multimodais. Esta seção discute algumas abordagens presentes na literatura relevantes ao contexto deste trabalho. Primeiramente, são apresentados alguns trabalhos que utilizam métodos de combinação simples, seguido por exemplos de trabalhos que realizam a fusão baseando-se em grafos, seguidos de trabalhos baseados na agregação por re-ranqueamento e finalizando com trabalhos que utilizam redes convolucionais baseadas em grafos.

Dentre os muitos trabalhos relevantes no contexto de recuperação multimodal, está o realizado por Snoek, Worring e Smeulders (2005), que apresenta uma comparação de métodos de fusão precoce e fusão tardia no domínio de vídeos, considerando características visuais, de áudio e de texto. No cenário de fusão precoce ocorre a concatenação dos vetores de características a fim de criar uma recuperação baseada unicamente em um vetor. O vetor é utilizada em um SVM(*Support Vector Machine*) para aprender uma representação semântica, enquanto que para a fusão tardia a combinação de características é realizada em

relação as representações aprendidas pela SVM para as diferentes modalidades. Ambas as estratégias apresentam ganhos significativos em relação a comparação com as modalidades isoladas, com a fusão tardia atingindo resultados ligeiramente superiores.

Song, Wang e Tian (2015) propõem o teste de três abordagens clássicas simples não supervisionadas de fusão no cenário multimodal. Nesse trabalho foram aplicadas sobre imagens, considerando três descritores diferentes e texto, mas poderiam ser generalizadas para outras mídias. Cada abordagem é definida por uma regra de fusão para definição de uma nova distância, sendo de modo breve baseadas no princípio de máximo, mínimo e média. Mesmo com a aplicação de estratégias tão simples de fusão os ganhos foram significativos, principalmente para a modalidade da regra de máximo. Esses resultados são indicadores que apontam para os os benefícios da da combinação de múltiplas modalidades de características de um objeto.

Dentre as abordagens utilizando de modelos profundos temos trabalhos como o de Han et al. (2017) que desenvolveram técnicas de recuperação de video baseadas em elementos de texto como elementos de consulta, caracterizando um caso de *Cross-modal Retrieval*, em que o elemento de consulta é de um tipo diferente do obtido pela recuperação. Nestas técnicas em específico tal processo ocorre utilizando de um modelo *Fast Fisher Vector Products*, onde utilizando do elemento de busca se obtém imagens fracamente rotuladas da internet para treino. As imagens obtidas com esse processo são uma representação do elemento de busca, e então cada video é tratado como um conjunto de grupos desordenados de imagens no banco de dados, representado por um único vetor Fisher, construído através da saída de uma rede neural convolucional e utilizado para recuperação de informações.

Quanto a abordagens baseadas em grafos, há o trabalho de Wang et al. (2012) que propõe uma abordagem de recuperação multimodal baseada no re-ranqueamento utilizando de grafos. Utilizando múltiplas modalidades de características visuais são construídos múltiplos grafos de distância para a coleção de dados, e a partir de pontuações de relevância e peso das modalidades é criado um novo ranqueamento de semelhança sendo uma representação da combinação dos grafos todas as modalidades.

Também utilizando de grafos, Dourado, Tabbone e Torres (2019) utilizam de um método para combinação de características de múltiplas modalidades para imagem, onde são empregados grafos de agregação de ranqueamento para realizar a fusão segundo o método proposto em (DOURADO; PEDRONETTE; TORRES, 2019). O método trata-se de uma abordagem que visa gerar uma representação dos múltiplos ranqueamentos que permita a fácil comparação de modo global entre elementos e uma representação de correlação em relação a toda a coleção de dados. Em virtude da natureza dessa abordagem, a recuperação proposta está atrelada a medidas de similaridade entre os grafos de fusão. Em uma abordagem conceitualmente similar, mas diferente em aspectos de construção, Ah-

Pine, Csurka e Clinchant (2015) propõem a fusão de características visuais e textuais para recuperação multimídia utilizando grafos. Ah-Pine se baseia em duas principais estratégias para combinação de características, uma que analisa as similaridades *cross-media*, analisando os objetos mais próximos das instâncias em múltiplas modalidades, e combinando linearmente a pontuação obtida pelo processo. Sua segunda abordagem é baseada em caminhos aleatórias pelo grafo, determinadas pela matriz estocástica, uma representação que contém a probabilidade de ir de um estado ao outro, baseando-se nos vetores de característica e no contexto de múltiplas modalidades do conjunto de dados, gerando assim novos ranqueamentos baseados no caminho percorrido.

Também utilizando grafos, Dourado, Tabbone e Torres (2020) propõem um modelo de aprendizado baseados em grafos de fusão de ranqueamento para recuperação de informação multimodal. De modo que os ranqueamentos de diferentes modalidades obtidos para uma consulta têm suas relações descritas em um grafo, que possui a habilidade de representar relações entre modalidades, instâncias e da coleção como um todo.

Quanto a trabalhos que utilizam de aprendizado não supervisionado sem o uso de grafos, temos abordagens como o de Li et al. (2014), que propõem um *framework* de agregação de ranqueamentos para um sistema que utiliza de descritores das modalidades visuais e de texto e realiza a tarefa de *multimedia geocoding*, onde unem ambos os ranqueamentos aplicando métodos tradicionais de agregação de listas como Borda Count (MERLIN, 2003) e métodos como o *Reciprocal Rank Fusion*(RRF) (CORMACK; CLARKE; BUETTCHER, 2009), que de modo resumido ordena os objetos de acordo com uma score dentre as referências reciprocas nas listas ranqueadas.Esse mesmo trabalho também considera uma abordagem multiplicativa proposta pelos autores, baseada na multiplicação das diferentes pontuações de similaridade obtidas pelas listas ranqueadas de modo que tal multiplicação leva a pontuações altas obtidas de uma característica a serem propagadas para outras.

Em uma abordagem baseada em ranqueamentos não supervisionada Mourão, Martins e Magalhães (2014) combinam características de texto e imagem para recuperação multimodal no campo médico. Combinando tanto elementos de recuperação por pontuação quanto por ranqueamento propõem a abordagem *Inverse Square Rank*, que combina a abordagem de ranqueamento inverso com o método RRF.

Considerando abordagens que utilizam Redes Convolucionais baseadas em Grafos (*Graph Convolutional Networks* - GCN), temos trabalhos como o de Misraa et al. (2020), que utiliza de GCNs para obter uma recuperação Multimodal de imagens, de modo que as informações visuais e de conceitos(textuais) são modeladas em forma de um grafo, criando relações entre elas, visando capturar a rica informação que a vizinhança de cada imagem fornece em relação aos conceitos que ela está atrelada. Tal grafo serve então de entrada para uma rede neural de grafo, que tem como meta produzir para cada nó vetores de

características que serão utilizados no processo de recuperação.

Também utilizando GCNs, Gu et al. (2021) realizam um sistema de recuperação multimodal baseado em grafos de aprendizado profundo para recuperarem imagens médicas, mais especificamente exames de Endomicroscopia. A multimodalidade desse projeto se dá ao combinar informações dos exames de imagem de Endomicroscopia com dados de histologia desses exames, fazendo com isso com que o modelo profundo aprenda a partir desses dados relacionar informações cruzando ambos e gerando vetores de características que são depois utilizados para recuperação de imagens.

Shen et al. (2020) propõem uma abordagem nomeada GCNs para *hashing* semi-supervisionada. O objetivo consiste em, a partir da modalidade de imagem e de texto, realizar recuperação. A abordagem procede de modo que duas GCNs, uma para Imagem e uma para Texto são treinadas com suas saídas conectadas a redes siamesas, a cada instância visual alimentada na rede é alimentada junto com uma de texto que descreve a imagem, com cada qual indo para a respectiva GCN, sobre a saída das redes siamesas é calculada um código Hash que geram os as representações (*embenddings*) para a recuperação.

Zeng et al. (2021) propõem para recuperação de momentos de vídeos uma abordagem que utiliza de informações visuais e textuais com aprendizado de grafo, de modo que são construídos dois grafos, um de relações visuais e outro de textuais e após o treinamento sobre ambos é extraída a representação de grafo para cada modalidade que é utilizada na recuperação por *matching* nos grafos entre modalidades gerados.

2.5 Redes Convolucionais Baseadas em Grafos

O recente sucesso de redes neurais artificiais gerou significativo interesse na pesquisa de reconhecimento de padrões e *data mining*. Incluindo com isso tarefas como detecção de objetos, reconhecimento de ação, reconhecimento de voz, etc. Porém os espaço Euclidiano apresenta consideráveis limitações para aplicações que requerem modelagem de relações em arranjos mais complexos. Nestes cenários os problemas são mais facilmente representados em forma de grafos, dado que trata-se de uma estrutura consideravelmente complexa, podendo apresentar formatos irregulares, com diferentes tamanhos e nós com diferentes números de ligações. Tais características tornam operações como convoluções consideravelmente mais complexas no domínio de grafos. Além disso, dados representados em estruturas de grafos acabam por derrubar a premissa de grande parte de algoritmos de aprendizado de máquina de que as instâncias de um conjunto de dados são independentes, dado que em um espaço de grafo podem possuir diversas relações com outros componentes da coleção. (WU et al., 2021).

Os primeiros registros de GNNs(*Graph Neural Network*) datam de 1997, quando Sperduti e Starita (1997) aplicaram redes neurais de grafo a grafos acíclicos. Porém somente

diversos anos depois o conceito de GNN foi delineado, por Gori, Monfardini e Scarselli (2005) e posteriormente aperfeiçoado por Scarselli et al. (2009) e Gallicchio e Micheli (2010).

Porém, o conceito de redes de grafo convolucionais apesar de ter tido seus primeiros estudos em 2009, por Micheli (2009), somente se tornou um tópico de considerável interesse de pesquisa mais recentemente, quando diversas abordagens de GCNs surgiram considerando tendo como objetivo, levar o notável sucesso de resultados de CNNs para problemas representados em domínio de grafo (ATWOOD; TOWSLEY, 2015; NIEPERT; AHMED; KUTZKOV, 2016; GILMER et al., 2017; KIPF; WELLING, 2016).

De modo geral, GCNs generalizam a operação de convolução de uma malha para dados em formato de grafo. A principal ideia é gerar uma representação do nó w por meio de agregar suas próprias *features* x_w e as *features* de nós vizinhos x_v em que v pertença a vizinhança de w ($v \in \mathcal{N}(w)$). Tal processo é realizado de modo que cada nó de realiza tal agregação com sua vizinhança relevante e repassa esses valores para a próxima camada, como ilustrado na Figura 6. Existem duas principais categorias de GCNs, levando em conta a estrutura de suas convoluções, sendo elas espectrais e espaciais. Redes convolucionais baseadas em grafos espectrais definem convoluções a partir de filtros, em uma perspectiva de processamento de sinal grafo, partindo da teoria espectral de grafo para definir o que é relevante em meio as informações da vizinhança. Enquanto redes convolucionais baseadas em grafos espaciais abordam convoluções de grafo por meio da agregação direta de informações dos vizinhos mais próximos para cada nó.

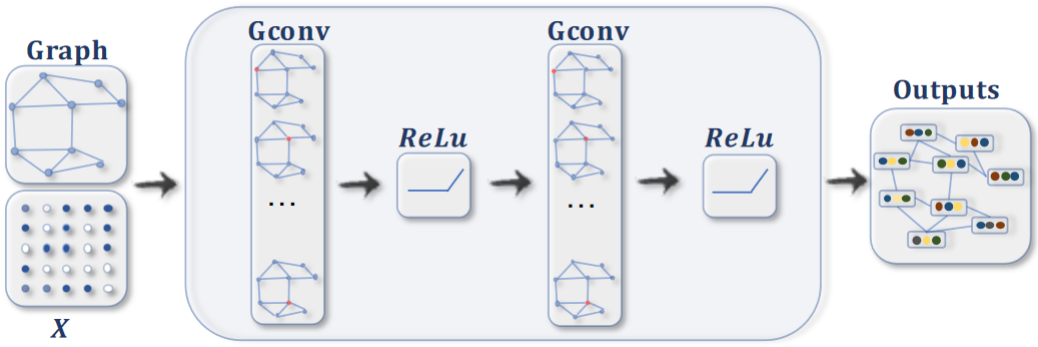


Figura 6 – Uma GCN com múltiplas camadas convolucionais, para cada camada cada nó encapsula a informação de sua vizinhança, agregando as outras informações a sua própria.Figura retirada de (WU et al., 2021)

.

Em geral, um modelo GCN pode ser descrito como uma função sobre uma matriz de *features* $\mathbf{X}$ e uma matriz de adjacência $\mathbf{A}$, de modo que: $\mathbf{Z} = f_{gcn}(\mathbf{X}, \mathbf{A})$. Tal matriz de adjacência define a vizinhança inicial do conjunto de dados que será alimentado na GCN. A arquitetura de GCN utilizada nesse projeto trata-se da descrita por Kipf et. al em (KIPF; WELLING, 2016). Sendo uma GCN espectral, que realiza convoluções por

meio de uma aproximação de primeira ordem de filtros espectrais localizados de grafo, visando uma otimização de custo computacional que ainda assim apresente resultados consistentes. Desse modo as convoluções são limitadas a *kernels* cujo espectro é uma função afim de autovalores. Podendo ser definida como a multiplicação ($\star$) de um sinal $x \in \mathbb{R}^N$ (um escalar para cada nó) com um filtro $g_\theta = \text{diag}(\theta)$ parametrizado por $\theta \in \mathbb{R}^N$ no domínio de Fourrier, de modo que:

$$g_\theta \star x = \mathbf{U} g_\theta \mathbf{U}^\top x \tag{2.2}$$

Em que $\mathbf{U}$ é a matriz de autovetores do grafo Laplaciano normalizado $L = \mathbf{I_N} - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, com uma matriz diagonal de seus autovalores $\mathbf{\Lambda}$ e $\mathbf{U}^\top x$ sendo o grafo da transformada de Fourier de x, $\mathbf{I_N}$ sendo uma matriz identidade e $\mathbf{D}$ por sua vez a matriz de graus do grafo. Podemos então entender g_θ como uma função dos autovalores de L, tal que $g_\theta(\mathbf{\Lambda})$. Avaliando a Equação 2.2, é notável que ela é computacionalmente custosa, dado que a multiplicação da matriz de autovetores $\mathbf{U}$ já resultaria em uma complexidade $\mathcal{O}\left(N^2\right)$. Além disso computar a autodecomposição de L teria um custo proibitivo para grafos de grande escala de componentes. Desse modo para contornar tal complexidade, foi apresentado por Hammond, Vandergheynst e Gribonval (2009) que $g_\theta(\Lambda)$ poderia ser aproximada com precisão suficiente pela expansão truncada em termos dos polinômios de Chebyshev $T_k(x)$ até a ordem K:

$$g_{\prime}(\Lambda) \approx \sum_{k=0}^{K} \theta'_k T_k(\tilde{\Lambda}) \tag{2.3}$$

Considerando uma $\mathbf{\Lambda}$ reescalada, tal que $\tilde{\mathbf{\Lambda}} = \frac{2}{\lambda_{\max}}\Lambda - I_N \cdot \lambda_{\max}$, tal que denota o maior autovalor de $L \cdot \theta' \in \mathbb{R}^K$, em que λ trata-se de um fator de peso, tem-se agora um vetor dos coeficientes de Chebyshev. Recursivamente, os polinômios de Chebyshev podem ser definidos como $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$, com $T_0(x) = 1$ e $T_1(x) = x$.

Voltando a definição de convolução como um sinal de x com um filtro $g_{\theta\prime}$, agora temos:

$$g_{\theta\prime} \star x \approx \sum_{k=0}^{K} \theta'_k T_k(\tilde{L})x \tag{2.4}$$

Com $\bar{L} = \frac{2}{\lambda_{\max}}L - I_N$; como pode ser verificado, levando em conta $\left(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\right)^k = \mathbf{U}\mathbf{\Lambda}^k\mathbf{U}^\top$. Note que essa expressão é agora K-orientada, dado que é um polinômio de ordem K, no Laplaciano. Em outras palavras, depende apenas dos nós que estão no máximo a k passos de distância do nó central (uma vizinhança de ordem K). Importante ressaltar que com essa aproximação, realizada na Equação 2.4, a complexidade diminui para $\mathcal{O}(|\mathcal{E}|)$, linear em relação ao número de arestas.

3 Materiais, Métodos e Protocolo Experimental

Esse Capítulo tem como objetivo apresentar os materiais, métodos e protocolo experimental utilizados para cada abordagem. A Seção 3.1 apresenta os descritores utilizados para extração de características de imagem, vídeo e áudio utilizadas pelas abordagens propostas. Considerando os vetores de características e estruturas de ranqueamento obtidas, foram aplicados os métodos de *manifold learning* descritos na Seção 3.2, de acordo com a pipeline de cada abordagem. Além disso, a Seção 3.3 apresenta sobre o algoritmo de treinamento utilizado para treinar GCNs de modo completamente não supervisionado, conforme estratégia utilizada nas abordagens de recuperação de vídeos propostas. Em seguida, a Seção 3.4 define os conjuntos de dados utilizados para a avaliação experimental de cada cenário. Por fim, a Seção 3.5 discute sobre as métricas de eficácia utilizadas para avaliação de ambos os cenários de abordagens.

3.1 Extração de Características

Esta seção tem como objetivo detalhar os seis modelos utilizados para a extração de características nesse projeto, abordando detalhes de arquitetura e treinamento de cada modelo utilizado. Dois desse modelos (a Rede Residual Para Reconhecimento de Imagem e a Rede Residual I3D para reconhecimento de Ações em vídeos) além de serem utilizados nas abordagens de vídeo também foram utilizados na recuperação multimodal baseada em aprendizado de representações de imagens (apresentada no Capítulo 5).

3.1.1 Descritores de Imagem

Para a modalidade de imagem nas abordagens de recuperação de vídeo, foi extraído de cada vídeo o quadro mais representativo utilizando do histograma médio de cores. De modo que consideramos todos os quadros do vídeo foi realizada o cálculo de um histograma médio e o quadro que mais se aproximasse desse histograma foi selecionado, pois esse quadro seria o que melhor sumarizaria informações que mais aparecem no vídeo. Então sobre o quadro escolhido, foram extraídas as características de imagem.

Então, sobre o quadro mais representativo foram extraídas as características de imagem.

- **Rede Residual Para Reconhecimento de Imagem (HE et al., 2015)**

O primeiro modelo pré-treinado utilizado para a extração de características da

modalidade de imagem é uma Rede Residual (ResNet), proposta por He et al. (2016). Este modelo possui 18 camadas internas e foi treinado para a tarefa de classificação no conjunto de dados ImageNet (DENG et al., 2009).

A arquitetura ResNet é um caso especial da arquitetura CNN, que popularizou a ideia de "ignorar conexões"(*skip connections*), também conhecido como conexões de atalho. De acordo com He et al. (2016), com o aumento da profundidade da rede, a acurácia torna-se saturada e então degrada rapidamente. Para resolver esse problema, eles usaram blocos residuais, cuja ideia subjacente é incluir um mecanismo de atalho entre cada duas camadas da rede, adicionando a entrada diretamente à saída. Desta forma, quanto mais camadas, menor é a mudança de cada camada para a entrada, tornando as redes residuais fáceis de otimizar e obter maior precisão quando a profundidade da rede aumenta, produzindo resultados melhores do que as redes não residuais.

Este modelo foi escolhido devido aos seus excelentes resultados em muitas tarefas em competições de classificação. Uma rede semelhante implementada pelos mesmos autores, mas com maior profundidade, ficou em primeiro lugar em detecção ImageNet, localização ImageNet, detecção COCO e segmentação COCO nas competições ILSVRC e COCO 2015.

- **Rede Convolucional Densamente Conectada Para Reconhecimento de Imagem (HUANG; LIU; WEINBERGER, 2016)**

O segundo modelo selecionao para extração de características de imagens trata-se de uma DenseNet de 121 camadas, também treinada sobre o conjunto de dados ImageNet (DENG et al., 2009) por Huang, Liu e Weinberger (2016), para a tarefa de classificação e imagem.

Estudos recentes comprovaram que redes CNNs podem ser mais profundas, sem apresentar problemas de saturação de acurácia se a arquitetura possuir conexões mais curtas entre camadas perto da entrada e próximas as saída (HUANG; LIU; WEINBERGER, 2016). Considerando esse principio, redes convolucionais densas (DenseNets) conectam cada camada a cada outra camada da rede, num princípio de *feed-foward*. Reduzindo com isso consideravelmente o problema de *vanishing* gradiente.

A escolha desse modelo se deu pelo interesse em comparar uma arquitetura ResNet com outra semelhante em termos de extração de características e também a performance que o modelo demonstrou na tarefa de classificação sobre a coleção ImageNet.

3.1.2 Descritores de Vídeo

- **Rede Residual I3D para Reconhecimento de Ações em Vídeos (MONFORT et al., 2019b)**

O primeiro modelo de extração de características de vídeo trata-se de uma arquitetura de rede neural I3D (Inflated 3D), uma ResNet(Residual Network) 3D, caracterizada

por ser uma versão que contém filtros aprendidos em 2D "inflados"para compreender a dimensão temporal. Tal estratégia é comumente adotada para aproveitar o forte ganho de aprendizado de características de se treinar uma rede em *datasets* notáveis de imagem, como por exemplo o ImageNet (DENG et al., 2009) anteriormente ao treinamento para vídeos, melhorando significativamente o desempenho da arquitetura. Essa melhora se dá principalmente pelo motivo que modelos 3D comumente contêm um número de parâmetros exponencialmente maior que modelos 2D o que os torna difíceis de treinar do zero.

Por tal, anteriormente ao treinamento no conjunto de dados *Moments in Time* (MONFORT et al., 2019a), conjunto de vídeos composto por um milhão de elementos rotulados, distribuídos entre 339 classes, foi realizado o pré-treinamento da arquitetura no conjunto *ImageNet*. Tal modelo foi escolhido por ser uma arquitetura idealizada pelos desenvolvedores da coleção *Moments in Time*, além dos ótimos resultados obtidos.

- **Rede Residual com Convoluções Mistas para Reconhecimento de Ações em Vídeos (TRAN et al., 2018)**

O segundo modelo de extração de características de vídeo trata-se de uma *ResNet Mixed Convolution*, em que o modelo é composto tanto de camadas com convoluções 3D quanto convoluções 2D. O fundamento que apoia tal arquitetura está no fato de que, enquanto as primeiras camadas processam e aprendem informações de movimento por meio de convoluções 3D, as camadas próximas ao fim aprendem informações espaciais utilizando de convoluções 2D, e da informação extraída por meio das convoluções 3D. Essa abordagem conseguiu apresentar desempenho semelhante a seu modelo equivalente apenas com Convoluções 3D para a tarefa de classificação que foi projetada, apresentando 3 vezes menos parâmetros internos, tornando tanto o treinamento consideravelmente mais rápido como também a extração de características.

O modelo em questão possui 18 camadas internas e foi treinado sobre o conjunto de dados Kinects (KAY et al., 2017), um conjunto de dados de grande escala, com 400 classes de ações diferentes. A escolha desse modelo se deu pela performance a nível de estado da arte quando em comparação com outros modelos sobre a tarefa de classificação sobre a coleção de dados Kinects.

3.1.3 Descritores de Áudio

- **Rede Convolucional para Entendimento de Eventos em Áudio (KONG et al., 2020)**

O primeiro modelo utilizado para a extração de características de áudio dos conjuntos de dados trata-se de um dos modelos propostos e treinados por Kong et al. (2020), esta arquitetura em específico é uma rede neural convolucional a qual recebe como entrada uma representação conjunta, composta da união do espectrograma de Mel com o

Wavegram, uma representação criada pelos autores que se assemelha ao espectrograma de Mel porém é aprendido com intermédio de uma rede neural, visando aprender e gerar uma representação da frequência rítmica do áudio.

Esse modelo foi treinado sobre o *dataset* AudioSet (GEMMEKE et al., 2017), um conjunto de cerca de 2.1 milhões de instâncias de áudio dividas entre 527 classes, sobre esse conjunto de dados a arquitetura apresentou MAP consideravelmente alto em comparação ao apresentado na literatura como estado da arte, o que levou o mesmo a ser escolhido como extrator de características de áudio.

- **Rede Residual Para Reconhecimento de Áudio (CHEN et al., 2020)**

O segundo modelo de áudio trata-se de um modelo de arquitetura baseada em ResNets com 18 camadas internas, implementado por Chen et al. (2020) e treinado sobre o conjunto de dados VGG-Sound (CHEN et al., 2020), um conjunto de grande escala com mais de 200 mil clipes de áudio divididos em 309 classes variadas. O modelo recebe uma representação dos áudios que trata-se da transformada de Fourier de curto tempo sobre o formato original, que resulta em um espectrograma de dimensão 257×500.

O motivo que levou a escolha desse modelo foi o fato de seu treinamento ter ocorrido sobre um conjunto de dados diferente do outro modelo de extração de características de áudio e também pela acurácia significativa que a arquitetura apresentou sobre o conjunto de dados VGGSound e também sobre o conjunto AudioSet (GEMMEKE et al., 2017).

3.2 Métodos de Manifold Learning

Esta seção discute os métodos de aprendizado não supervisionado baseados em *manifold learning* utilizados para fusão de múltiplas modalidades tanto na abordagem de recuperação de vídeos quanto de imagens. Além disso, essa seção também descreve brevemente o arcabouço UDLF, que contém a implementação dos métodos.

3.2.1 LHRR

O primeiro método, denominado LHRR (*Log-based Hypergraph of Ranking References*) (PEDRONETTE et al., 2019) usa um modelo de hipergrafo para explorar as informações de similaridade e transformá-las em modelos de ranqueamento. Os grafos são comumente representados por conjuntos de vértices (nós) e suas conexões correspondentes (arestas ou links). Os hipergrafos, por sua vez, são uma generalização desses grafos que permitem a conexão de qualquer número de vértices e a representação de relações de similaridade de ordem superior.

Com base nas referências dos ranqueamentos, é construída a representação dos hipergrafos. Para construir uma representação contextual de amostras de dados, é usada a

abordagem de hiperarestas. Seguindo uma função baseada em log, os pesos são atribuídos aos objetos em cada hiperaresta. Por meio dele, é possível explorar as informações de similaridade codificadas. Tal similaridade é obtida através do resultado do produto das similaridades com suas respectivas hiperarestas. O objetivo, então, é obter uma função de similaridade mais eficaz. A ideia desse novo conjunto de ranqueamentos calculado e da função de similaridade obtida é utiliza-los para aprimorar a eficacia do conjunto de rankings final. O método LHRR é utilizado em tarefas de ranqueamento não supervisionado, a fim de melhorar a eficácia da recuperação dos resultados, uma vez que é capaz de identificar relações de similaridade mais confiáveis e capturar a estrutura geométrica dos conjuntos de dados. O método também pode ser usado para tarefas de fusão de ranqueamentos, que é o objetivo com que usamos o método neste trabalho.

3.2.2 BFS-Tree

Usando uma estrutura de árvore, o algoritmo BFS-Tree de aprendizado não supervisionado (*Breadth-First Search Tree of Ranking References*) (Carlos Guimarães Pedronette; VALEM; TORRES, 2021) é aplicado para explorar as informações de similaridade codificadas nas referências das listas ranqueadas. A fim de obter os resultados dos top-k do ranqueamento, o método fornece uma representação hierárquica dos resultados de ranqueamento, codificando as relações de vizinhança de primeira e segunda ordem obtidas por meio de referências entre as listas ranqueadas. Calculado com base nas medidas de correlação de ranqueamento, os pesos das arestas atribuídos aos elementos da árvore representam a similaridade.

O BFS-Tree também é utilizado para descobrir relacionamentos de similaridade subjacentes, de modo que os elementos da árvore são representados com base em seu caminho até a raiz e seus respectivos pesos, assim entre as folhas, novas conexões são estabelecidas. Essas conexões tornam possível descobrir novas relações de semelhança. Uma estrutura em árvore também permite, além de novas conexões de similaridade, analisar a frequência dos elementos na árvore. Normalmente, uma indicação sólida de similaridade pode ser obtida pela coocorrência de elementos em diferentes níveis da estrutura da árvore, enquanto uma baixa ocorrência pode ser uma indicação de ruído. Graças à consideração de informações de similaridade extraídas de todas as árvores construídas, é possível computar uma medida de similaridade mais global e efetiva entre pares do conjunto de dados, aplicando o BFS-Tree tanto para melhorar a eficácia das estruturas de ranqueamento, quanto para fusão dos mesmos.

3.2.3 Arcabouço UDLF

Desenvolvido por Valem et al., o *UDLF*(*Unsupervised Distance Learning Framework*) (VALEM; PEDRONETTE, 2017) trata-se de um arcabouço que tem como objetivo o

acesso direto à utilização e avaliação de diversos métodos de aprendizado não supervisionado. Essa ferramenta define um modelo geral que possibilita, a partir de arquivos de configuração, o uso de diferentes métodos baseados em informações de listas ranqueadas. O projeto está disponível publicamente[1] sob os termos da licença GPLv2 de tal forma que a comunidade científica pode acessar, utilizar e compartilhar mudanças.

O arquivo de configurações (ilustrado na Figura 7) possui os parâmetros que são necessários para selecionar o método e configura-lo antes de executa-lo. Este arquivo composto das seguintes seções:

- *ConFigurações Gerais:*possibilita a escolha do método e da tarefa;
- *Dados de Entrada:* formato e localização dos arquivos de entrada;
- *Dados de Saída:* formato e localização dos arquivos de saída;
- *Avaliação:* métricas de eficácia a serem calculadas;
- *Parâmetros:* parâmetros do método selecionado.

[1] <http://www.ic.unicamp.br/~dcarlos/UDLF/>

```
#CATEGORY 1. GENERAL CONFIGURATION
UDL_TASK = UDL #(UDL|FUSION): Selection of task to be executed
UDL_METHOD = CPRR #(NONE|CPRR|RLRECOM|RLSIM|CONTEXTRR|RECKNNGRAPH|RKGRAPH|
CORGRAPH): Selection of method to be executed

#CATEGORY 2. INPUT FILE SETTINGS
SIZE_DATASET = 1400 #(TUint): Number of images in the dataset
INPUT_FILE_FORMAT = MATRIX #(MATRIX|RK): Format of input file
INPUT_MATRIX_TYPE = DIST #(DIST|SIM): Type of matrix file
INPUT_RK_FORMAT = NUM #(NUM|STR): Format of ranked list file
MATRIX_TO_RK_SORTING = HEAP #(HEAP|INSERTION): Convert matrix to rks
NUM_INPUT_FUSION_FILES = 2 #(TUint): Number of files for FUSION tasks
INPUT_FILES_FUSION_1 = input1.txt #Path of the first input file
INPUT_FILES_FUSION_2 = input2.txt #Path of the second input file
#INPUT_FILES_FUSION_* = input*.txt #Path of the *th input file
INPUT_FILE = input.txt #Path of the main input file (matrix/rks)
INPUT_FILE_LIST = list.txt #Path of the list file
INPUT_FILE_CLASSES = classes.txt #Path of the classes file
INPUT_IMAGES_PATH = images/ #Dataset images path (required only for visual results)

#CATEGORY 3. OUTPUT FILE SETTINGS
OUTPUT_FILE = TRUE #(TBool): Generate output file(s)
OUTPUT_FILE_FORMAT = MATRIX #(RK|MATRIX): Format of output file
OUTPUT_MATRIX_TYPE = DIST #(DIST|SIM): Type of matrix file to output
OUTPUT_RK_FORMAT = ALL #(NUM|STR|HTML|ALL): Output format for rks
OUTPUT_FILE_PATH = output #Path of the output file(s)
OUTPUT_HTML_RK_PER_FILE = 100 #(TUint): Number of rks for each html file
OUTPUT_HTML_RK_SIZE = 40 #(TUint): Number of images per ranked list
OUTPUT_HTML_RK_COLORS = TRUE #(TBool): Color borders around images
OUTPUT_HTML_RK_BEFORE_AFTER = TRUE #(TBool): Comparison of rks

#CATEGORY 4. EVALUATION SETTINGS
EFFICIENCY_EVAL = TRUE #(TBool): Enable efficiency evaluation
EFFECTIVENESS_EVAL = TRUE #(TBool): Enable effectiveness evaluation
EFFECTIVENESS_COMPUTE_PRECISIONS = TRUE #(TBool): Compute precisions
EFFECTIVENESS_COMPUTE_MAP = TRUE #(TBool): Compute MAP
EFFECTIVENESS_COMPUTE_RECALL = TRUE #(TBool): Compute recall
EFFECTIVENESS_RECALL_AT = 40 #(TUint): Position to compute recall
EFFECTIVENESS_PRECISIONS_TO_COMPUTE = 5, 20 #(TUint ["," TUint]*):
Precisions to be computed (unsigned integers separated by commas)

#CATEGORY 5. METHOD PARAMETERS
PARAM_RECKNNGRAPH_L = 200 #(TUint): Size of ranked lists
PARAM_RECKNNGRAPH_K = 15 #(TUint): Number of nearest neighbors
PARAM_RECKNNGRAPH_EPSILON = 0.0125 #(TFloat): Value used for the convergence criteria
```

Figura 7 – Arquivo de configuração do framework ULDF. Retirada de (VALEM; PEDRONETTE, 2017)

3.3 Deep Graph Infomax

O treinamento dos modelos de GNCs utilizados, foi realizado com o algoritmo de treinamento não-supervisionado *Deep Graph Infomax* (DGI) (VELIčKOVIć et al., 2019). Partindo de um cenário sem rótulos, o algoritmo define o que deve ou não estar conectado no grafo por meio de uma estratégia baseado em maximizar a informação local mútua, criando representações que capturam informações globais do grafo. Algo interessante de tal algoritmo, é que diferente da maioria dos método de treinamento não supervisionado de GCNs encontrados atualmente na literatura, o DGI não se baseia nem utiliza de caminhadas aleatórias (VELIčKOVIć et al., 2019).

Tal algoritmo assume uma configuração de aprendizado de grafo não supervisionado da forma descrita a seguir. Tem-se um conjunto de *features* de nós $\mathbf{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$, onde N trata-se do número de nós e $\vec{x}_i \in \mathbb{R}^F$ representa as *features* do nó i. Também se tem como dado de entrada a informação de relação entre esses grafos, podendo ser uma matriz de adjacência $\mathbf{A} \in \mathbb{R}^{N \times N}$. Em nosso trabalho utilizamos uma matriz de adjacência

sem pesos na aresta, sendo apenas $\mathbf{A}_{ij} = 1$ se há conexão $i \to j$ no grafo, ou $\mathbf{A}_{ij} = 0$ caso não haja. Partindo de tais informações, o objetivo do algoritmo é encontrar um *encoder*, $\mathcal{E} : \mathbb{R}^{N\times F} \times \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times F'}$ e $\mathcal{E}(\mathbf{X}, \mathbf{A}) = \mathbf{H} = \left\{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_N\right\}$ sejam representações de alto nível de $\vec{h}_i \in \mathbb{R}^{F'}$ para cada nó i, em outras palavras, sendo a forma de coleção da matriz de *embenddings* $\mathbf{Z}$. Essas representações contendo informações não só do nó como de toda sua vizinhança e do conjunto de dados como um todo, podem então ser utilizadas para diversas tarefas, desde classificação, recuperação de informação, etc.

Deste modo, a abordagem para aprender o *enconder* se baseia em maximizar a informação local mútua, visando obter uma representação de nó (local) que capture informações globais de todo o grafo, representado por um vetor sumário $\vec{s}$. Para obter vetores sumário a nível de grafo, o algoritmo utiliza de uma função de leitura $\mathcal{R} : \mathbb{R}^{N\times F} \to \mathbb{R}^F$, e utiliza para sintetizar as representações obtidas em nível de fragmentos para o nível de grafo, $\vec{s} = \mathcal{R}(\mathcal{E}(\mathbf{X}, \mathbf{A}))$. Como uma *proxy* por maximizar a informação mútua local, o DGI aplica um descriminador, $\mathcal{D} : \mathbb{R}^F \times \mathbb{R}^F \to \mathbb{R}$, tal que $\mathcal{D}\left(\vec{h}_i, \vec{s}\right)$ representa as pontuações(*scores*) de probabilidade atribuídas para esse par de sumários de fragmentos(de modo que são mais altas para fragmentos contidos no sumário).

Amostras negativas para $\mathcal{D}$ são então providas emparelhando o sumário $\vec{s}$ de $(\mathbf{X}, \mathbf{A})$ com representações de fragmentos $\vec{h}_j$ de um grafo alternativo, $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{A}})$. Em um cenário multi-grafo, tais grafo são obtidos como outros elementos de um conjunto de treinamento. Para um cenário de apenas um grafo, é utilizada uma explicita função de corrupção, $\mathcal{C} : \mathbb{R}^{N\times F} \times \mathbb{R}^{N\times N} \to \mathbb{R}^{M\times F} \times \mathbb{R}^{M\times M}$, sendo ela necessária para se obter exemplos de amostras negativas a partir do grafo original, de modo que $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{A}}) = \mathcal{C}(\mathbf{X}, \mathbf{A})$. A escolha do procedimento de amostragem negativa irá reger os tipos específicos de informação estrutural que se deseja capturar como subproduto desta maximização.

Por padrão, para tal tarefa, o algoritmo utiliza de princípios apresentados por Hjelm et al. em (HJELM et al., 2019), utilizando de uma função objetiva do tipo ruído-contrastante, com um função de perda de entropia cruzada padrão(BCE) entre as amostras positivas e o produto dos marginais(exemplos negativos). Tal função pode ser definida por:

$$\mathcal{L} = \frac{1}{N+M}\left(\sum_{i=1}^{N} \mathbb{E}_{(\mathbf{X},\mathbf{A})}\left[\log \mathcal{D}\left(\vec{h}_i, \vec{s}\right)\right] + \sum_{j=1}^{M} \mathbb{E}_{(\widetilde{\mathbf{X}},\widetilde{\mathbf{A}})}\left[\log\left(1 - \mathcal{D}\left(\overrightarrow{\widetilde{h}}_j, \vec{s}\right)\right)\right]\right) \tag{3.1}$$

Em que M trata-se do número de amostras marginais(negativas) e $\mathbb{E}_{(\mathbf{X},\mathbf{A})}$ o valor esperado considerando as matrizes$\mathbf{X}$ e $\mathbf{A}$. Tal abordagem, efetivamente maximiza a informação mútua entre $\vec{h}_i$ e $\vec{s}$, baseado na divergência de Jensen-Shannon (HJELM et al., 2019).

Como todas as representações derivadas dos fragmentos do grafo são voltadas para preservar a informação mútua do sumário global do grafo, tal algoritmo permite descobrir novas informações globais e ainda assim preservar similaridades em nível de fragmentos locais.

Para um cenário de apenas um grafo, como o utilizado nesse projeto, o algoritmo do DGI pode ser resumido nos seguintes passos:

1. Criar amostras negativas usando a função de corrupção: $(\tilde{\mathbf{X}}, \widetilde{\mathbf{A}}) \sim \mathcal{C}(\mathbf{X}, \mathbf{A})$.

2. Obter representações locais de fragmentos $\vec{h}_i$, para o grafo de entrada passando pelo *encoder*: $\mathbf{H} = \mathcal{E}(\mathbf{X}, \mathbf{A}) = \left\{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_N\right\}$

3.Obter representações de fragmentos das amostras negativas, passando também pelo *encoder*: $\widetilde{\mathbf{H}} = \mathcal{E}(\widetilde{\mathbf{X}}, \widetilde{\mathbf{A}}) = \left\{\vec{h}_1, \vec{h}_2, \ldots, \overrightarrow{\tilde{h}}_M\right\}$

4. Criar o sumário do grafo passando suas representações de fragmentos locais pela função de leitura mencionada anteriormente: $\vec{s} = \mathcal{R}(\mathbf{H})$.

5. Atualizar os parâmetros de $\mathcal{E}, \mathcal{R}$ e do descriminador $\mathcal{D}$ aplicando gradiente descendente para maximizar a Equação 3.1.

3.4 Conjuntos de Dados

Essa seção apresenta todos os conjuntos de dados utilizados na avaliação experimental das abordagens propostas.

3.4.1 Conjuntos de Dados para Recuperação Multimodal de Vídeos

Para a avaliação experimental das abordagens propostas para recuperação multimodal de vídeos foram foram utilizados três conjuntos de dados de vídeos, os quais são descritos a seguir:

• **UCF-101 (SOOMRO; ZAMIR; SHAH, 2012):**

O primeiro conjunto de dados utilizado trata-se do UCF-101 *Action Recognition Data Set* (SOOMRO; ZAMIR; SHAH, 2012), um conjunto de dados para reconhecimento de ações que possui 13.320 vídeos divididos em 101 classes. As classes compreendem interações humanas com objetos, movimentação humana, interações entre humanos, humanos tocando instrumentos musicais e esportes. Seus vídeos possuem cerca de 3 a 8 segundos, apresentam apenas uma ação e são fruto de cenários não controlados e extraídos da internet, tendo variações de angulo, iluminação, câmera de filmagem, pose e plano de fundo, sendo realistas e consideravelmente complexos, representando com isso uma tarefa significativamente desafiadora para a recuperação de informação. Do conjunto de dados, 6.640 vídeos possuem trilha de áudio, cerca de metade da coleção, sendo particularmente adequado para essa abordagem multimodal. Na Figura 8 são apresentados alguns exemplos de vídeos presentes no UCF-101.

Figura 8 – Exemplos de instâncias presentes no conjunto de dados UCF101.

• **MSR-VTT (XU et al., 2016):**

O segundo conjunto de dados utilizado trata-se do MSR-VTT (XU et al., 2016), um conjunto de dados com 10 mil vídeos, distribuídos em 20 categorias (música, pessoas, jogos, esportes, notícias, educação, shows de tv, cinema, animações, veículos, tutoriais, viagem, ciência, animais, infantil, comida, cozinhando, documentário, beleza e anúncios). Ainda que comumente empregado para descrição de vídeo, pela disparidades entre categorias dos conjuntos de dados torna-se interessante o emprego desse conjunto de dados para recuperação multimídia multimodal, ainda mais considerando que todos as instâncias do conjunto de dados possuem trilha de áudio. Na Figura 9 estão dispostos exemplos de instâncias contidas na coleção MSR-VTT.

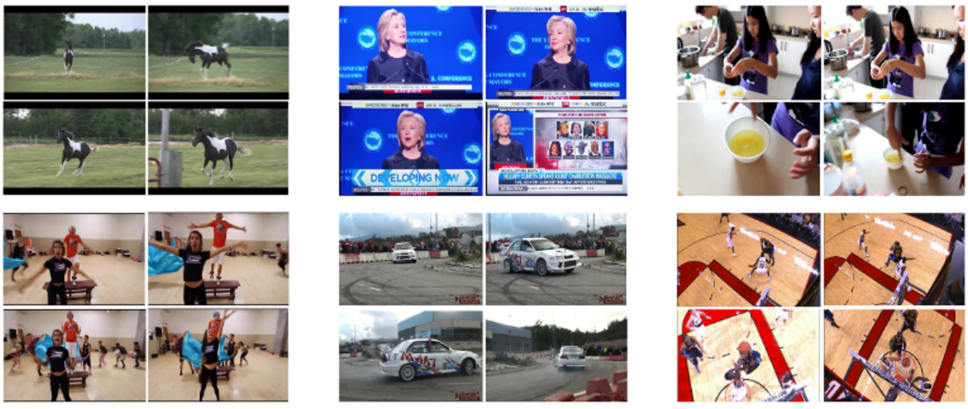


Figura 9 – Exemplos de instâncias presentes no conjunto de dados MSR-VTT.

• **TV Human Interaction Dataset (PATRON-PEREZ et al., 2012):**

O terceiro e último conjunto de dados utilizado para essa parte é o TV *Human Interaction Dataset* (PATRON-PEREZ et al., 2012), o qual trata-se de um conjunto de menor escala, possuindo 300 vídeos divididos em cinco classes (apertos de mão, *high*

fives, abraços, beijos e negativos). Seus vídeos foram retirados de programas de televisão, tendo com isso uma iluminação, visibilidade e ângulos de câmera favoráveis a facilitar a legibilidade das ações. A Figura 10 apresenta exemplos de vídeos contidos no conjunto.

Figura 10 – Exemplos de instâncias presentes no conjunto de dados TV Human Interaction Dataset.

3.4.2 Conjuntos de Dados para Recuperação Multimodal de Imagens

Para a experimentação da abordagem de recuperação de imagens baseada em *manifold learning e representation learning* foram utilizados três conjuntos de dados públicos de imagens, descritos a seguir:

- **Willow *Actions***:

O Willow *Actions* (DELAITRE; LAPTEV; SIVIC, 2010) é composto por 911 imagens estáticas distribuídas em sete classes de ações. Suas imagens foram extraídas do Flickr e possuem apenas uma das sete ações (Interagir com Computador, Fotografar, Tocar Instrumento, Andar de Bicicleta, Andar a Cavalo, Correr e Andar). Em geral, possuem um plano de fundo simples, sem muitos elementos além da ação. Na Figura 11 há uma amostra de imagens do conjunto de dados Willow *Actions*.

- **Ikizler Dataset**

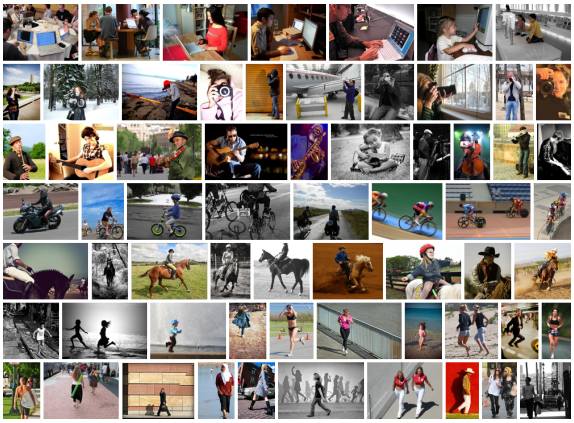


Figura 11 – Amostra de imagens do Willow *Actions* (DELAITRE; LAPTEV; SIVIC, 2010).

O segundo conjunto de dados denominado Ikizler Dataset (TANISIK; ZALLUHOGLU; IKIZLER-CINBIS, 2016), é uma coleção de 1972 imagens, divididas em dez classes (boxe, jantar, aperto de mão, *highfive*, abraço, chute, beijo, festa, fala e conversa), onde cada classe tem pelo menos 150 imagens. Este conjunto de dados é consideravelmente mais complexo do que o Willow *Actions*, pois possui classes com ações muito semelhantes, como aperto de mão e *highfive*. Na Figura 12 está disposta uma amostra de imagens do conjunto.

Figura 12 – Amostra de imagens do conjunto de dados Ikizler *Dataset*. (TANISIK; ZALLUHOGLU; IKIZLER-CINBIS, 2016).

• **Stanford 40 *Actions Dataset***

Finalmente, o terceiro conjunto de dados foi o Stanford 40 *Actions Dataset* (YAO et al., 2011), o qual é composto por 9532 imagens e possui 40 classes de ações diferentes com pelo menos 180 imagens para cada categoria. Devido ao maior número de classes, em comparação com os demais conjuntos de dados, torna-se um desafio ainda mais significativo

obter uma recuperação eficaz nesse conjunto de dados. As imagens que formam o conjunto de dados foram extraídas do Bing, Google e Flickr. A Figura 13 apresenta uma amostra do conjunto de dados.

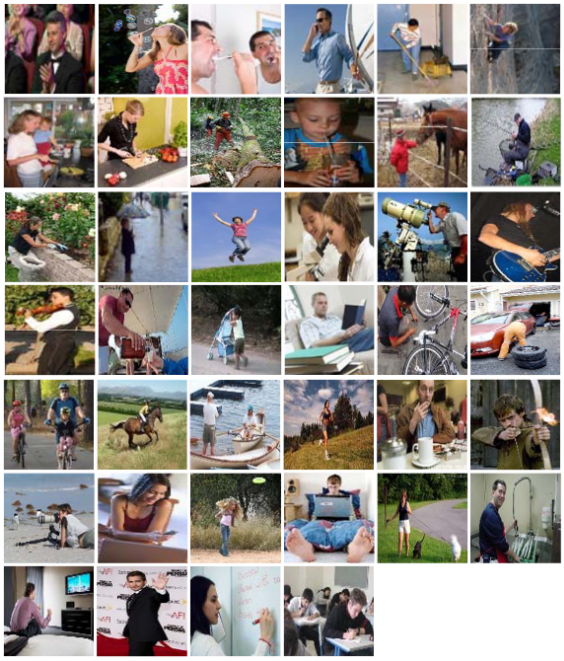


Figura 13 – Amostra de imagens do conjunto de dados Stanford 40 Actions (YAO et al., 2011).

3.5 Métricas de Eficácia

Há diversas métricas de eficácia disponíveis na literatura propostas com o objetivo de avaliar quantitativamente os resultados de tarefas de recuperação de informação. No contexto deste trabalho, para avaliação de eficácia das abordagens de ambos os cenários, foram utilizadas as duas seguintes métricas:

• **Precisão ($P@H$):** trata-se da média de acertos considerando os h primeiros elementos de todas as listas ranqueadas do conjunto. Seja N o número de objetos na coleção, H o números de posições no topo de cada lista ranqueadas que estão sendo consideradas e t o número de verdadeiros positivos (objetos que pertencem a mesma classe daquele a qual a lista ranqueada pertence) nessas H posições, temos:

$$P(N) = \sum_{j=1}^{N} \left(\frac{t_j}{H} \right). \tag{3.2}$$

• **MAP (*Mean Average Precision*):** Trata-se da média da precisão para todos os H contemplados pelo conjunto das listas ranqueadas. Nessa métrica é analisada a quantidade de elementos de mesma classe do objeto de consulta que estão nas posições

próximas ao topo de sua *ranked list*, em outras palavras nas posições mais significativas dela. A métrica MAP pode ser definida como: sendo q um elemento de consulta, N_r o número de objetos significativos á consulta e $(r_i|i = 1, 2, ..., d)$ um vetor de instâncias relevantes, vetor o qual possui tamanho d e onde r indica o valor numérico da relevância do item que ocupa a i-ésima posição, com 0 caso não significante e 1 caso contrário, definimos MAP como:

$$AP = \frac{1}{N_r} \sum_{i=1}^{d} \left(\frac{r_i}{i} \sum_{j=1}^{i} r \right). \tag{3.3}$$

4 Recuperação Multimodal de Vídeos Baseada em Manifold Learning e GCNs

Este capítulo descreve as três abordagens propostas para recuperação multimodal de vídeos baseadas em métodos de *manifold ranking* e GCNs. O restante do capítulo está organizado de modo que a Seção 4.1 apresenta uma visão geral da motivação de tais abordagens, enquanto a Seção 4.2 apresenta a primeira abordagem, que utiliza apenas de métodos de *manifold learning*. Por sua vez, as Seções 4.3 e 4.4 apresentam respectivamente as abordagens que utilizam de métodos de *manifold learning* com GCNs em estratégias de fusão primária e tardia, denominadas *Early* GCN e *Late* GCN. Por fim, a Seção 4.5 apresenta sobre a avaliação experimental realizada considerando os conjuntos apresentados anteriormente e discute os resultados obtidos. É interessante observar que nesse trabalho para a parte de vídeos consideramos a multimodalidade como as diferentes modalidades de informação que as características se baseiam (áudio, vídeo e imagem).

4.1 Visão geral

Haja vista a abundância e diversidade de informações disponíveis em dados de conteúdo multimídia, um dos objetivos deste trabalho consiste em explorar a complementariedade de dados de diferentes modalidades em tarefas de recuperação. Pretende-se assim avaliar o potencial de combinação de modalidades para aumentar a eficácia dos resultados. Com estes propósitos, consideramos uma tarefa de recuperação multimodal de vídeos em um cenário não supervisionado. Assim como descrito anteriormente, foram utilizadas diversos descritores de características para diferentes modalidades. A combinação foi realizada por uma estratégia de fusão tardia, utilizando métodos de *manifold learning* baseados em ranqueamento e também GCNs, resultando em três abordagens com diferentes *pipelines*, sendo a primeira apenas com métodos de *manifold ranking*, e duas outras mais complexas incorporando tanto *manifold ranking* quanto GCNs.

Assim como mencionado anteriormente o fundamento por trás da combinação de métodos de *manifold ranking* com GCNs está em combinar a capacidade de métodos *manifold ranking* de gerar medidas de distância que considerem informações de toda a coleção, com a habilidade de redes GCNs de criarem representações (*embenddings*) que considerem não apenas a informações do próprio nó como também de sua vizinhança.

4.2 Recuperação Baseada em Manifold Learning

A primeira abordagem utilizando apenas métodos de *manifold ranking* é ilustrada na Figura 14. O fluxo dessa abordagem inicia-se na extração das características, de modo que tendo uma função de extração de características f para cada descritor m, aplicamos f_m sobre cada instância de cada conjunto de dados $\mathcal{X}$, tal que $f_m(\mathcal{X}) = \mathbf{X}_m$, sendo $\mathbf{X}$ a matriz de *features* extraídas para a modalidade m. Desse modo partindo das m matrizes de características, calculamos um ranking inicial para cada descritor utilizando da distância Euclidiana, gerando com isso m conjuntos de listas ranqueadas $\mathcal{T}_m$. Tais rankings são utilizados como dados de entrada para os métodos de *manifold ranking fusion*, que combinam as informações desses ranqueamentos visando obter um ranqueamento final com melhor eficácia $\mathcal{T}_n^*$, a partir de uma função g_S, definida na Seção 2.1.2 , tal que:

$$\mathcal{T}_n^* = g_S(\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_m) \tag{4.1}$$

É válido ressaltar que a Figura 14 apresenta uma simplificação visual para esse processo, dado que para cada modalidade foram utilizados dois descritores, resultando em seis ranqueamentos de entrada no algoritmo de *manifold ranking fusion*, conforme detalhado na Seção 3.1.

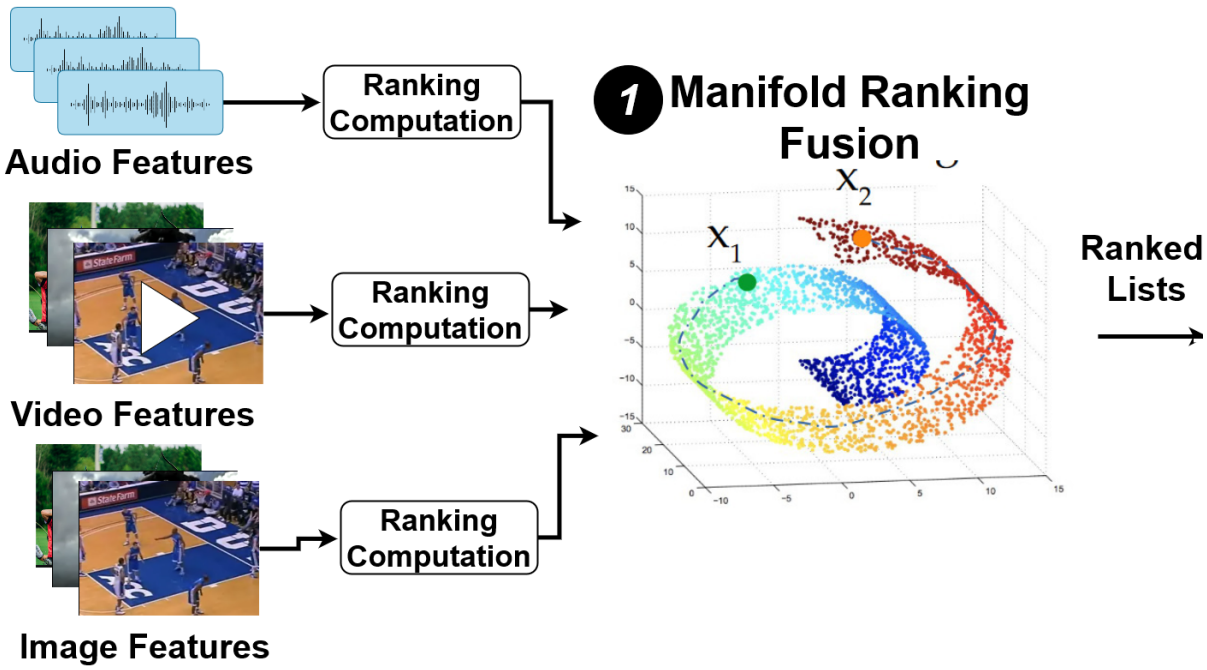


Figura 14 – Fluxo Geral da Abordagem apenas utilizando Manifold Ranking.

4.3 Recuperação Baseada em Early GCN

A Figura 15 apresenta nossa segunda abordagem proposta, Early GCN. Essa abordagem utiliza dos resultados obtidos pelo ranqueamento final da abordagem anterior($\mathcal{T}_n^*$) para criação de um grafo k-NN recíproco (Passo **I**). O grafo utilizado pode ser definido seguindo a notação utilizada em (PEDRONETTE; GONçALVES; GUILHERME, 2018c), como um grafo não direcionado $G_r = (V, E)$, em que o conjunto de vértices V é dado pela coleção de dados $V = \mathbf{X}$ e cada objeto da coleção é representado por um nó. A coleção de arestas E é calculada baseando-se nos k-vizinhos mais próximos recíprocos.

Deste modo, para determinar a vizinhança recíproca, nós primeiro definimos o conjunto de vizinhança. Dado um objeto de consulta q, um conjunto de vizinhança $\mathcal{N}(q,k)$, que contém os k objetos mais similares ao objeto q pode ser definido como

$$\begin{aligned}\mathcal{N}(q,k) = \{\mathcal{S} \subseteq \mathbf{X}, |\mathcal{S}| = k \wedge \forall obj_i \in \mathcal{S} \\ obj_j \in \mathbf{X} - \mathcal{S} : \tau_q(i) < \tau_q(j)\}\end{aligned}$$

Então dado que as relações de vizinhos mais próximos não é simétrica o conjunto dos k-vizinhos mais próximos recíprocos de um objeto q pode ser definido como

$$\mathcal{N}_r(q,k) = \{obj_i \in \mathcal{N}(q,k) \wedge obj_q \in \mathcal{N}(i,k)\}.$$

Enquanto que o conjunto de arestas E pode ser formalmente definido como:

$$E = \{(obj_q, obj_j) \mid obj_j \in \mathcal{N}_r(q,k)\}$$

Assim sendo, nós podemos interpretar que irá ter uma aresta de obj_q para obj_j se os objetos forem vizinhos recíprocos até as k posições. Tal grafo $G_r(V,E)$ é utilizado para criar a matriz de adjacências do conjunto de dados, como temos um cenário que utilizamos de um grafo sem pesos nas arestas, temos uma matriz A de nXn tal que $A_{ij} = 1$ se houver conexão entre os objetos i e j, ou zero caso contrário.

Além disso realizamos uma fusão precoce (concatenação) das *features* de todas as modalidades, que resulta em uma matriz $\mathbf{X}_c$, tal que $\mathbf{X}_c = \mathbf{X}_1 \oplus \mathbf{X}_2 \oplus ... \oplus \mathbf{X}_m$. Então, uma vez calculadas as matrizes $\mathbf{A}$ e $\mathbf{X}_c$, elas são utilizadas como dados de entrada para a GCN (Passo **II**). A GCN utilizada trata-se de uma GCN de uma camada, que resulta em uma saída de *embenddings* com dimensão (1,128) treinada pelo já descrito algoritmo DGI, por 100 épocas. A saída da GCN pode ser definida como o resultado de uma função que gera uma matriz de *embeddings* $\mathbf{Z}$, tal que: $\mathbf{Z} = f_{gcn}(\mathbf{X}_c, \mathbf{A})$. Em seguida, as representações contidas em $\mathbf{Z}$ são comparadas utilizando a distância Euclidiana para o cálculo de conjuntos de *ranked lists* (definidos pelo conjunto $\mathcal{T}$). A partir das estruturas de ranqueamento

calculadas, aplicamos métodos de *Manifold Learning* (Passo **III**) para obter medidas de distâncias mais eficazes e globais em um ranqueamento final $\mathcal{T}_n^*$, tal que $\mathcal{T}_n^* = g_S(\mathcal{T})$.

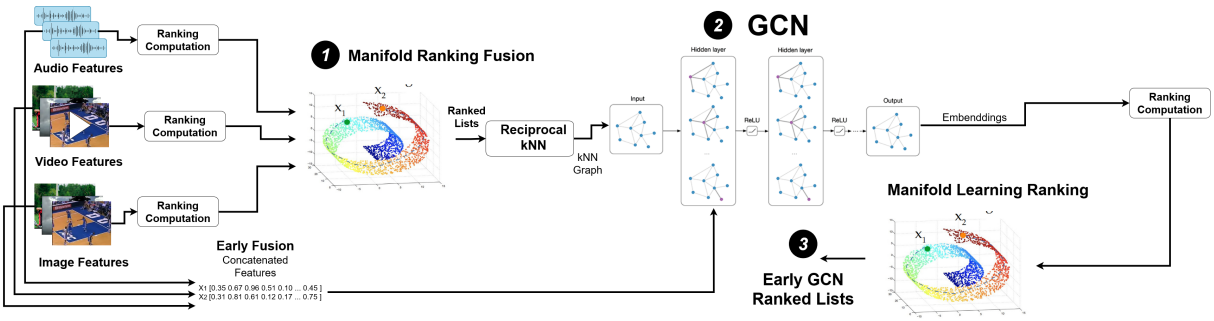


Figura 15 – Fluxo Geral da Abordagem Early GCN.

4.4 Recuperação baseada em Late GCN

A terceira abordagem proposta, denotada por *Late GCN*, está ilustrada na Figura 16. Em que a definição da matriz de adjacências $\mathbf{A}$, ocorre de mesmo modo que na abordagem *Early GCN*, baseando-se em métodos de *manifold ranking* (Passo **I**) e na criação de um grafo k-NN recíproco utilizando a mesma definição da abordagem Early GCN. Porém, a diferença mais relevante dentre essas abordagens está no uso de vários modelos GCN treinados utilizando o algoritmo DGI. Uma vez calculada a matriz de adjacências definida pelo grafo, a matriz de *features* de cada descritor m é utilizada como dados de entrada em uma GCN diferente, juntamente com a matriz de adjacência. Nessa abordagem, a saída de cada GCN resulta em um conjunto de *embenddigs*, resultando em 6 conjuntos $\mathbf{Z}_m$, tal que considerando cada modelo GCN independente como uma função $f_{gcn}m$, temos $\mathbf{Z}_m = f_{gcn_m}(\mathbf{X}_m, \mathbf{A})$ para cada um dos m descritores de características (Passo **II**). A partir dos conjuntos gerados, realizamos o cálculo de conjuntos de *ranked lists* com a distância Euclidiana para cada um deles, obtendo m ranqueamentos $\mathcal{T}_m$. Por fim, métodos de *manifold learning fusion* são utilizados para combinar os diferentes conjuntos, em um processo de fusão tardia para obter um ranqueamento final $\mathcal{T}_n^*$, tal que $\mathcal{T}_n^* = g_S(\mathcal{T}_1, ..., \mathcal{T}_m)$ (Passo **III**).

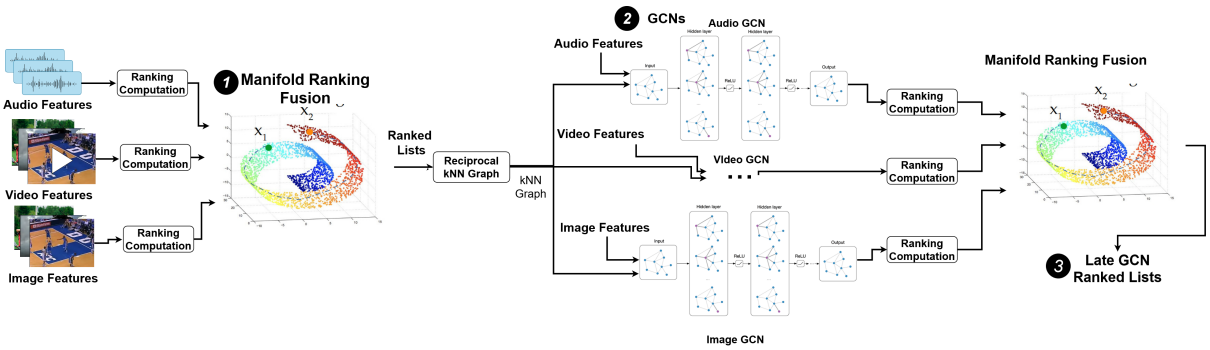


Figura 16 – Fluxo Geral da Abordagem Late GCN.

4.5 Avaliação Experimental

Essa Seção descreve os detalhes da avaliação experimental realizada. Em que as três abordagens propostas foram avaliadas para cada um dos três conjuntos descritos anteriormente (UCF-101, MSR-VTT,TV *Human Interaction*) e para cada um dos dois métodos de *manifold learning*(LHRR e BFS-Tree), também descritos anteriormente.

As abordagens foram comparadas com *baselines* clássicas de agregação de ranqueamento e também com métodos mais recentes. As *baselines* utilizadas foram Borda Count (MERLIN, 2003), CombSUM (ZHOU; DEPEURSINGE; MULLER, 2010), Regra do Produto (ZHOU; DEPEURSINGE; MULLER, 2010) e a mais recente *Graph Fusion* (ZHANG et al., 2012).

Para melhor comparar o impacto da combinação de informações de diferentes modalidades, foi também realizada a avaliação da eficácia de recuperação dos conjuntos de *ranked lists* de cada descritor isoladamente. De modo que a partir da distância Euclidiana, definimos o conjunto de *ranked lists* de cada descritor.

4.5.1 Resultados

Para o conjunto de dados TV *Human Interaction Dataset* os resultados estão dispostos na Tabela 1, de modo que são apresentadas primeiramente os resultados das métricas Precisão e MAP para cada conjunto de *features* isolados e em seguida, são apresentados os resultados das *baselines*, seguidos pelas abordagens propostas.

É possível notar que a abordagem *Early* GCN com o método BFS-Tree supera tanto os descritores individualmente quanto *baselines* clássicas para a maioria das métricas (P@20, P@50, MAP), porém o mesmo não ocorre com as outras abordagens, chegando a ficar abaixo de alguns descritores isolados, assim como as *baselines*. Deste modo, temos que a *Early* GCN com o método BFS-Tree apresenta um ganho relativo de MAP de 7.01% em relação ao melhor descritor isolado(ResNet I3D), 29,37% em relação a melhor *baseline* (CombSUM) e 28,99% em relação a melhor abordagem sem presença de GCNs (LHRR).

Para o conjunto de dados MSR-VTT os resultados estão dispostos na Tabela 2. Dentre as abordagens, a *Late* GCN apresentou os melhores resultados de recuperação em relação a métrica MAP , com o método LHRR, a qual supera todas as *baselines* e descritores isolados, apresentando um ganho relativo de MAP de 6.12% em relação ao melhor descritor isolado (Video Resnet I3D), 31.59% em relação ao melhor resultado obtido por *baselines* (CombSUM) e 6.37% em relação a melhor abordagem proposta sem uso de GCNs (BFS-Tree).

Como é possível notar para o conjunto de dados UCF-101 (Tabela 3), todas as três abordagens propostas superam os resultados dos descritores isoladamente quando utilizado

Tabela 1 – Resultados obtidos sobre o conjunto de dados *TV Human Interaction*.

Método	**P@5**	**P@20**	**P@50**	**P@100**	**MAP**
Descritores Individualmente					
Áudio ResNet	0.4440	0.3090	0.2815	0.2605	0.2936
Áudio CNN	0.4167	0.2815	0.2513	0.2360	0.2711
Imagem ResNet	0.4460	0.3158	0.2685	0.2457	0.2834
Imagem DenseNet	0.4580	0.3145	0.2706	0.2463	0.2865
Vídeo Resnet I3D	**0.5720**	0.4392	0.3660	**0.3114**	0.3820
Vídeo Resnet MC	0.4160	0.2860	0.2509	0.2405	0.2724
Baselines					
Borda Count	0.5073	0.3638	0.3079	0.2719	0.3127
CombSUM	0.5087	0.3697	0.3083	0.2684	0.3157
Regra do Produto	0.5100	0.3607	0.3077	0.2702	0.3139
Graph Fusion	0.5613	0.3467	0.2912	0.2469	0.2965
Abordagem Proposta					
Manifold Ranking LHRR	0.4860	0.3622	0.3072	0.2768	0.3169
Manifold Ranking BFS-Tree	0.5300	0.3688	0.2977	0.2615	0.3158
Early GCN + LHRR	0.3727	0.2538	0.2099	0.1891	0.3560
Early GCN + BFS-Tree	0.4613	**0.4508**	**0.3873**	0.2962	**0.4088**
Late GCN+LHRR	0.4507	0.3478	0.2956	0.2622	0.3066
Late GCN+BFS-Tree	0.4467	0.3245	0.2920	0.2537	0.2938

do método de *Manifold Ranking* LHRR. Assim como superam as baselines utilizadas, tanto as clássicas quanto a recente para todas as métricas com exceção da P@5, em que o método CombSUM apresenta o melhor resultado. Por fim, é importante ressaltar que a abordagem sem uso de GCNs se destacou em relação as outras, ao contrário do que ocorre para todos os outros conjuntos de dados. A abordagem apresenta ganhos relativos de MAP de 34.49% em relação a melhor *baseline* (CombSUM) e 41.09% em relação ao melhor descritor individualmente (ResNet Vídeo I3D) .

De modo geral, dado o comportamento das abordagens compostas por GCNs em relação ao conjuntos de dados é possível notar que a abordagem *Early* GCN atingiu resultados mais eficazes em conjuntos menores, enquanto a *Late* GCN em conjuntos maiores. É importante ressaltar que com exceção do conjunto UCF-101, as abordagens que compreendem GCNs via de regra demonstraram melhor eficácia de recuperação do que a composta apenas por *Manifold Ranking*. Como é possível analisar com nossos resultados, nossas abordagens via de regra apresentaram resultados inferiores para a métrica $P@5$ em comparação com as *baselines* e descritores isolados, tal comportamento se dá graças a criação de representações mais globais, o que por consequência prejudicam a informação local, afetando com isso métricas que se baseiam significativamente em informação local mas melhorando resultados de métricas que dependem mais de informações globais (MAP e P@100).

Visando analisar o comportamento de cada abordagem foi criada uma visualização

Tabela 2 – Resultados obtidos sobre o conjunto de dados MSR-VTT.

Método	**P@5**	**P@20**	**P@50**	**P@100**	**MAP**
Descritores Individualmente					
Áudio ResNet	0.2767	0.1384	0.1073	0.0952	0.0753
Áudio CNN	0.3387	0.2044	0.1697	0.1514	0.0932
Imagem ResNet	0.4313	0.2982	0.2568	0.2284	0.1341
Imagem DenseNet	0.4252	0.2931	0.2530	0.2269	0.1381
Vídeo Resnet MC	0.2824	0.1371	0.1043	0.0913	0.0726
Vídeo Resnet I3D	0.4756	0.3545	0.3120	0.2831	0.1731
Baselines					
Borda Count	0.4443	0.2858	0.2333	0.2039	0.1205
CombSUM	0.4947	0.3365	0.2800	0.2453	0.1396
Regra do Produto	0.4119	0.2770	0.2278	0.1973	0.1166
Graph Fusion	**0.5231**	0.2631	0.2434	0.1817	0.1384
Abordagem Proposta					
Manifold Ranking LHRR	0.4913	**0.3634**	**0.3193**	**0.2886**	0.1695
Manifold Ranking BFS-Tree	0.4440	0.2852	0.2233	0.1864	0.1057
Early GCN + LHRR	0.2992	0.1549	0.1149	0.0975	0.0708
Early GCN + BFS-Tree	0.3063	0.1548	0.1021	0.0962	0.0663
Late GCN+LHRR	0.4472	0.3377	0.3041	0.2812	**0.1837**
Late GCN+BFS-Tree	0.4545	0.3319	0.1246	0.2743	0.1246

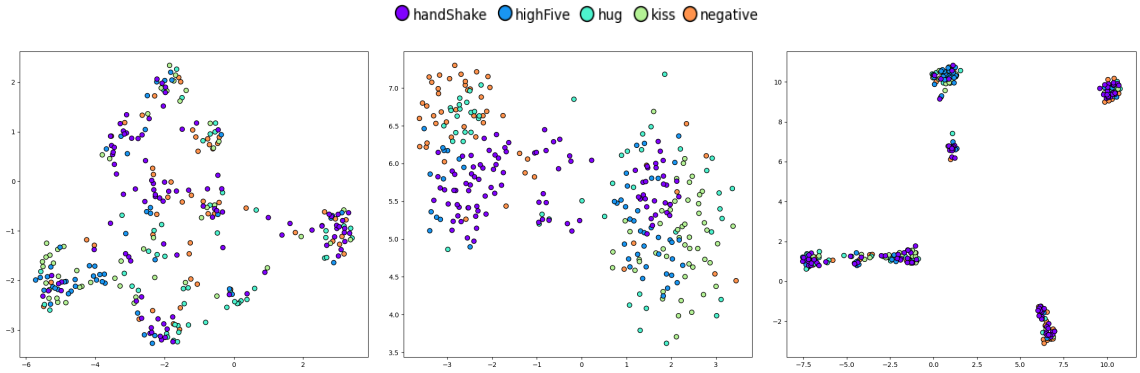
por meio do método de redução de dimensionalidade UMAP (MCINNES et al., 2018) sobre o conjunto TV Human Interactions, com o método que apresentou melhores resultados (BFS-Tree). Tal representação está disposta na Figura 17.

Nessa representação, é possível notar que os cenários que possuem GCNs se diferenciam bastante do cenário apenas com *Manifold Ranking*. De modo que a abordagem Early GCN apresenta a criação de grupos mais coesos, criando cada grupo com considerável distância entre seus componentes, apresentando uma separabilidade de classe superior a abordagem sem GCNs. Enquanto que para a abordagem Late GCN, a criação de grupos consideravelmente isolados para cada classe é notável, com uma separabilidade ainda maior do que a Early GCN.

Válido ressaltar que apesar da maior separabilidade de classes e grupos mais coesos, é observável que o fato de as abordagens forçarem uma maior separabilidade e coesão dos grupos, acarreta em significativos casos de instâncias sendo agrupadas com classes erradas. De modo que como trabalhos futuros seria interessante explorar modos de diminuir a coesão de tais grupos, ou aprimorar sua eficácia.

Tabela 3 – Resultados Obtidos sobre o conjunto UCF-101.

Método	P@5	P@20	P@50	P@100	MAP
Descritores Individualmente					
Áudio ResNet	0.2984	0.1426	0.0990	0.0786	0.0656
Áudio CNN	0.3425	0.2045	0.1628	0.1347	0.1128
Imagem ResNet	0.8577	0.5463	0.3975	0.3011	0.2436
Imagem DenseNet	0.8304	0.5393	0.4011	0.3067	0.2518
Vídeo Resnet MC	0.4647	0.2353	0.1489	0.1051	0.0729
Vídeo Resnet I3D	0.8782	0.6930	0.5944	0.5088	0.4670
Baselines					
Borda Count	0.6129	0.5004	0.4176	0.3406	0.2974
CombSUM	**0.9426**	0.7623	0.6451	0.5363	0.4899
Regra do Produto	0.4803	0.3752	0.3120	0.2579	0.2341
Graph Fusion	0.8123	0.7442	0.4319	0.5931	0.4585
Abordagem Proposta					
Manifold Ranking LHRR	0.8953	**0.8039**	**0.7459**	0.6792	**0.6589**
Manifold Ranking BFS-Tree	0.9128	0.7188	0.5944	0.4911	0.4539
Early GCN + LHRR	0.8208	0.7530	0.6952	0.6077	0.5721
Early GCN + BFS-Tree	0.8730	0.7631	0.6754	0.5384	0.4732
Late GCN+LHRR	0.8197	0.7675	0.7371	**0.6894**	0.6577
Late GCN+BFS-Tree	0.8090	0.7075	0.6431	0.5739	0.5401



(a) BFS-Tree **(b)** Early GCN + BFS-Tree **(c)** Late GCN + BFS-Tree

Figura 17 – Resultados sobre o TV Human Interaction dataset usando o método de redução de dimensionalidade UMAP para vizualização.

5 Aprendizado de Representações Multimodal para Recuperação de Imagens

Este capítulo apresenta o método proposto recuperação de imagens por meio do uso de aprendizado de representações. O restante do capítulo está organizado do seguinte modo: a Seção 5.1 apresenta a visão geral da abordagem proposta, enquanto a Seção 5.2 apresenta a definição formal adotada para as características do modelo de CNN 2D e a Seção 5.3 apresenta definição análoga para o modelo de CNN 3D. A Seção 5.4 apresenta a combinação dos ranqueamentos de ambos os modelos. Por fim, a Seção 5.5 apresenta a avaliação experimental e discute os resultados obtidos.

5.1 Visão Geral

Apesar do significativo sucesso de CNNs em tarefas de reconhecimento de imagens, esses modelos exigem uma grande quantidade de dados para permitir um treinamento eficaz, de forma a evitar cenários de sobreajuste (*overffiting*). Visando melhorar a capacidade de generalização de tais redes em cenários em que não há abundância de dados de treinamento, diversas abordagens têm sido propostas. A título de exemplo, técnicas de aumento de dados (*data augmentation*) têm sido amplamente utilizadas, de forma a tornar o conjunto de treinamento mais abrangente e diversificado e assim evitar o sobreajuste. Com objetivo análogo, apresentamos uma proposta de aprendizado de representação explorando conceitos de multimodalidade para aprendizado de representações. A representação obtida é mais abrangente e portanto, mais eficaz para a tarefas de recuperação de imagens em cenários não supervisionados. Válido ressaltar que trata-se de um aprendizado de representações multimodal, não uma recuperação multimodal, dado que consideramos como diferentes modalidades informações de diferentes representações como diferentes modalidades, mas ambas as representações se tratam de informações visuais estáticas.

A Figura 18 ilustra os principais passos da abordagem proposta. Cada uma das etapas é identificada na Figura por números romanos e referenciada no texto a seguir. Primeiramente, utilizamos uma formulação de transferência de aprendizado baseada em uma rede CNN 2D, treinada em outro conjunto de dados de grande escala. Os vetores de características obtidos por meio da CNN 2D servem para a criação de um primeiro conjunto de listas ranqueadas (Passo **I**). Esse primeiro ranqueamento define sequencias de imagens que são usadas como dados de entrada para o modelo 3D como uma sequência de quadros, de modo que para cada lista ranqueada a sequencia é composta da imagem de consulta e os top-n vizinhos mais próximos (Passo **II**). A partir do vídeo gerado, são

extraídos novos vetores de características, a partir dos quais pode-se calcular um novo conjunto de listas ranqueadas. Por fim, os conjuntos de ranqueamentos da CNN 2D e da CNN 3D são combinados utilizando métodos de *manifold learning* (Passo **III**).

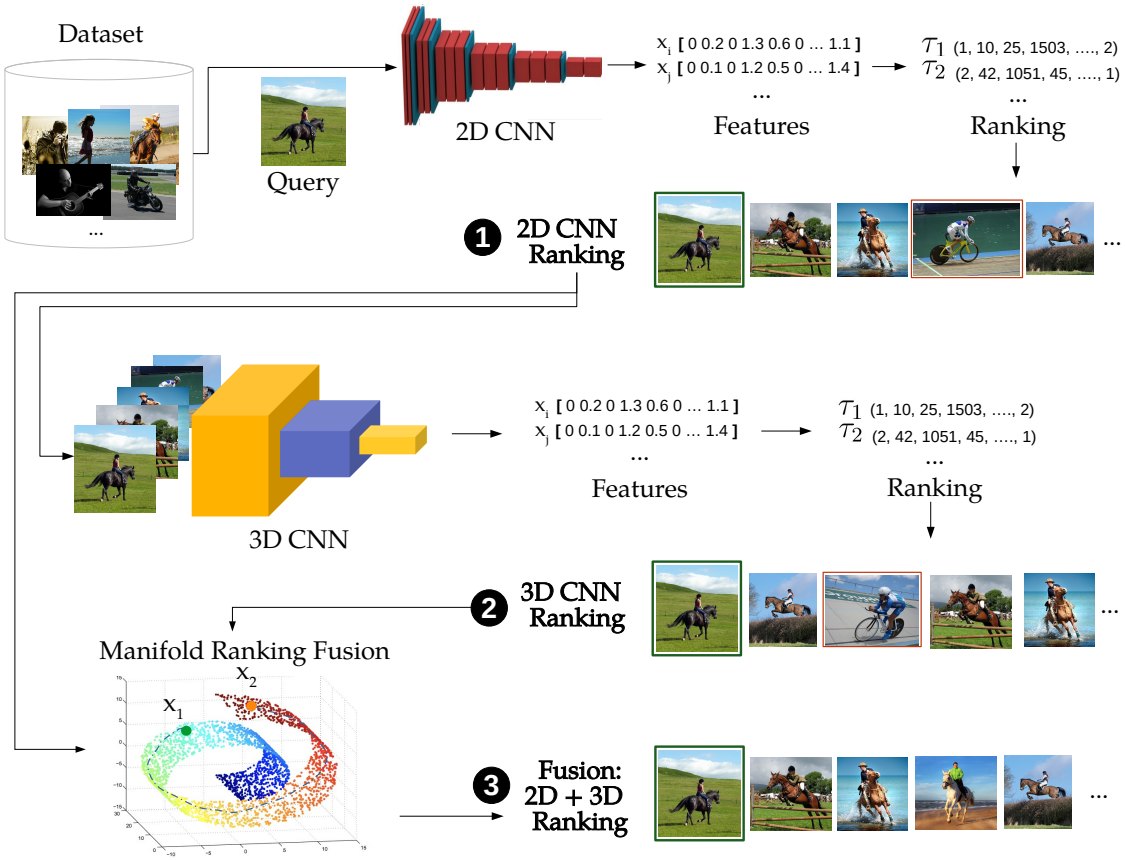


Figura 18 – Ilustração da abordagem de Aprendizado de Representações baseada em CNN 3D e *manifold learning*

.

Cada uma das etapas é formalmente definida nas seções seguintes. Os dois modelos utilizados para extração de características foram descritos na Seção 3.1. Os métodos de *manifold learning* utilizados estão descritos na Seção 3.2.

5.2 Representação baseada em CNN 2D

Esta seção apresenta a notação usada para recuperação de imagens e tarefas de ranqueamento, com base em trabalhos relacionados (TORRES; FALCÃO, 2006; Carlos Guimarães Pedronette; VALEM; TORRES, 2021) e define formalmente a representação da CNN 2D. Tal definição se aproxima bastante da realizada na Seção 2.1.1. Seja x uma imagem, ela pode ser definida formalmente por um par (D_x, I_x), em que:

- D_x é um conjunto finito de pontos (pixels) em $\mathbb{N}^2$, por exemplo, $D_x \subset \mathbb{N}^2$

- $I_x : D_x \rightarrow \mathbb{N}^3$ é uma função que atribui a cada pixel $p \in D_x$ um vetor $I(p) \in \mathbb{N}^3$ (uma cor no sistema RGB é atribuída a um pixel).

Seja $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ uma coleção de imagens, em que n denota o tamanho do coleção. As tarefas de ranqueamento e recuperação são realizadas com base nas características extraídas das imagens. Normalmente, CNNs 2D treinadas em conjuntos de dados de grande escala como ImageNet (DENG et al., 2009) são utilizados para extrair características para tarefas não supervisionadas por meio de transferência de aprendizagem. Deste modo a última camada totalmente conectada é frequentemente explorada para extração de características. Desta forma, um extrator de características pode ser formalmente definido como uma função f_2, em a notação subscrita se refere a uma característica da CNN 2D. Formalmente, a função $f_2 : \mathcal{X} \rightarrow \mathbb{R}^d$ calcula um vetor d -dimensional para uma dada imagem de coleção, tal que $\mathbf{v}_{2i} = f_2(x_i)$ e $\mathbf{v}_{2i} = [v_{2i1}, v_{2i2}, \ldots, v_{2id}]$.

Uma função de distância que calcula a distância entre duas imagens de acordo com a distância entre seus vetores de características correspondentes pode ser definida como ρ: $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$. Portanto, uma distância entre duas imagens x_i, x_j pode ser calculada por $\rho(\mathbf{v}_{2i}, \mathbf{v}_{2j})$. Uma tarefa geral de recuperação de imagem baseada em *features* extraídas pode ser modelada como o cálculo de uma lista ranqueada τ_{2q} em resposta a uma imagem de consulta x_q, de acordo com a função de distância ρ. Espera-se que as posições superiores das listas ranqueadas contenham as imagens mais relevantes em relação à imagem da consulta, de modo que o comprimento L das imagens ranqueadas são frequentemente considerados, com $L \ll n$. Também nos referimos a vizinhos como um pequeno conjunto de imagens semelhantes fornecidas pelas imagens com melhores classificações em k, de modo que $k \ll L \ll n$.

A lista de ranqueamento τ_{2q} pode ser definida como uma permutação (x_1, x_2, …, x_L) do subconjunto $\mathcal{X}_L \subset \mathcal{X}$, que contém as L imagens mais semelhantes para consultar a imagem x_q, tal que e $|\mathcal{X}_L| = L$. Formalmente, uma permutação τ_{2q} é uma bijeção do conjunto $\mathcal{X}_L$ no conjunto $[n_L] = \{1, 2, \ldots, L\}$. Para uma permutação τ_q, interpretamos $\tau_{2q}(x_i)$ como a posição (ou classificação) da imagem x_i na lista de classificação τ_{2q}. Se x_i é classificado antes de x_j na lista classificada de x_q (ou seja, se $\tau_{2q}(x_i) < \tau_{2q}(x_j)$), então ρ ($\mathbf{v}_{2q}$, $\mathbf{v}_{2i}$) $\leq$ ρ ($\mathbf{v}_{2q}$, $\mathbf{v}_{2j}$).

Tomando cada imagem $x_i \in \mathcal{X}$ como uma imagem de consulta x_q, um conjunto de listas ranqueadas $\mathcal{T}_2$ pode ser calculado, contendo uma lista ranqueada para cada imagem na coleção. Cada lista ranqueada estabelece uma relação de similaridade entre a imagem da consulta e todas as imagens na coleção $\mathcal{X}$. Portanto, o conjunto $\mathcal{T}_2$ codifica uma rica fonte de informações de similaridade / dissimilaridade sobre a coleção $\mathcal{X}$.

5.3 Representação baseada em CNN 3D

Enquanto nas CNNs 2D, as convoluções são aplicadas nos mapas de características 2D para calcular as características das dimensões espaciais (SANTOS; ALMEIDA, 2020), nas CNNs 3D são aplicadas para capturar as informações de movimento codificadas em vários quadros contíguos (SANTOS; SEBE; ALMEIDA, 2019). Em geral, as convoluções 3D são realizadas em estágios de CNNs para calcular características de dimensões espaciais e temporais (JI et al., 2013).

A informação da modalidade visual contida em um vídeo pode ser definida como uma sequência de imagens (ou quadros), de modo que $\sigma = (x_{t1}, x_{t2}, \ldots, x_{tm})$, onde o subscrito t_i denota a dimensão temporal e m denota o número de quadros no vídeo. CNNs 3D são frequentemente empregadas para extrair características de vídeos, por exemplo, representar as informações visuais codificadas em vídeo em uma representação vetorial dimensional de d que é usada para tarefas de recuperação e aprendizado de máquina.

Embora as sequências de imagens possam ser normalmente definidas por um vídeo, também podem ser definidas por uma lista ranqueada. Nossa hipótese considera que uma lista ranqueada computada em resposta a uma consulta de imagem pode codificar informações de similaridade contextual relevantes sobre a imagem, em substituição à dimensão temporal. Na verdade, muitas vezes fornece uma representação diversificada sobre a respectiva classe, uma vez que se espera que as posições superiores contenham diferentes imagens relevantes da mesma classe da imagem de consulta.

Uma lista ranqueada calculada com base na representação gerada pela CNN 2D é explorada para determinar a sequência. Tendo k_r denotando o tamanho da sequência definida pela lista ranqueada e seja $N_2(x_q, k_r)$ *oconjuntode* k_r mais semelhante a x_q de acordo com as características extraídas pela CNN 2D e listas ranqueadas, e definido como:

$$N_2(x_q, k_r) = \{\ \mathrm{C} \subseteq \mathcal{X}, |\mathcal{C}| = k_r \wedge \forall x_i \in \mathcal{C}, x_j \in \mathcal{X} - \mathcal{C} : \tau_{2q}(i) < \tau_{2q}(j)\}. \quad (5.1)$$

A sequência σ_q é uma permutação definida como uma bijeção do conjunto $N_2(x_q, k_r)$ *noconjunto* {1,2, ..., k_r}, que segue a ordem da lista ranqueada τ_{2q}. Se x_i é ranqueado antes de x_j na sequência σ_q (ou seja, se $\sigma_q(x_i) < \sigma_q(x_j)$), então $\tau_{2q}(x_i) \leq \tau_{2q}(x_j)$. Uma sequência pode ser definida para cada imagem $x_i \in \mathcal{X}$ a fim de computar um conjunto de sequências $\mathrm{S} = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$. Cada sequência pode ser analisada por uma CNN 3D com o objetivo de extrair novas características. Formalmente, uma CNN 3D pode ser definida como uma função $f_3 : \mathrm{S} \rightarrow \mathbb{R}^d$, que calcula um vetor d -dimensional para uma sequência de imagens, de modo que $\mathbf{v}_{3i} = f_3(\sigma_i)$ e $\mathbf{v}_{3i} = [v_{3i1}, v_{3i2}, \ldots, v_{3id}]$.

De forma análoga as características da CNN 2D, as listas ranqueadas podem ser calculadas com base nas características do modelo 3D. A lista ranqueada τ_{3q} também pode ser definida como uma bijeção do conjunto $\mathcal{X}_L$ para o conjunto $[n_L] = \{1, 2, \ldots, L\}$, de modo que se $\tau_{3q}(x_i) < \tau_{3q}(x_j)$, então $\rho\ (\mathbf{v}_{3q}, \mathbf{v}_{3i}) \leq \rho\ (\mathbf{v}_{3q}, \mathbf{v}_{3j})$.

Cada imagem $x_i \in \mathcal{X}$ também pode ser tomada como uma imagem de consulta x_q, a fim de obter um conjunto de listas ranqueadas $\mathcal{T}_3$. Desta forma, as características extraídas para cada imagem x_i por uma CNN 3D codificam não apenas informações espaciais, mas também informações de similaridade contextual de sua lista ranqueada τ_{2i}. Uma vez que as sequências processadas pelas CNNs 3D contêm uma representação mais diversificada para as imagens, o conjunto de listas ranqueadas $\mathcal{T}_3$, espera-se melhorar a generalização proporcionada pela representação e, consequentemente, melhorar a recuperação.

5.4 Fusão de Representações

Os conjuntos de listas ranqueadas obtidos a partir dos vetores de características dos modelos de CNN 2D e 3D codificam informações de similaridade relevantes e complementares. Enquanto o conjunto $\mathcal{T}_2$ fornece a informação de similaridade original e mais precisa, o conjunto $\mathcal{T}_3$ fornece uma representação de similaridade com mais diversidade. Portanto, essas informações podem ser combinadas a fim de obter uma medida de similaridade e ranqueamento mais eficazes.

Abordagens de *manifold learning* foram exploradas para melhorar e combinar o conjunto de listas ranqueadas (PEDRONETTE; GONçALVES; GUILHERME, 2018a; PEDRONETTE et al., 2019; Carlos Guimarães Pedronette; VALEM; TORRES, 2021).

O principal objetivo do método de aprendizagem *manifold* baseado em classificação é explorar as informações de similaridade codificadas no conjunto de listas ranqueadas, sendo capaz de capturar uma informação de similaridade global codificada na variedade do conjunto de dados. Com base nessa análise, um novo e mais eficaz conjunto de listas ranqueadas pode ser calculado, com o objetivo de melhorar a eficácia das tarefas de ranqueamento e recuperação. Considerando os dois conjuntos de listas ranqueadas $\mathcal{T}_2$ e $\mathcal{T}_3$, fornecidas pelas CNNs 2D e 3D, uma tarefa de fusão de ranqueamentos *manifold* pode ser definida como a função já apresentada g_S, tal que:

$$\mathcal{T}_f = g_S(\mathcal{T}_2, \mathcal{T}_3) \tag{5.2}$$

Espera-se que o conjunto $\mathcal{T}_f$ contenha listas ranqueadas mais eficazes que podem ser usadas em tarefas de recuperação.

5.5 Avaliação Experimental

Essa seção descreve a avaliação experimental empregada para a parte de aprendizado de representações, de modo que a Seção 5.5.1 apresenta e discute os resultados obtidos e por fim a Seção 5.5.2 apresenta visualizações dos resultados obtidos, por meio de métodos de redução de dimensionalidade.

Para cada instância de cada conjunto de dados foi originalmente gerada uma lista ranqueada a partir da distância Euclidiana dos vetores gerados pelo modelo de ResNet 2D. Utilizando esse ranqueamento inicial foram criados vídeos compostos da sequência dos top-n de cada lista ranqueada para a imagem de consulta, de modo que a sequencia era composta da imagem de consulta intercalada com seus vizinhos mais próximos(top-n). Utilizando dessa dinâmica foram criadas diferentes cenários de sequências, considerando 5, 10 e 15 de vizinhos mais próximos. Essas sequências por sua vez foram utilizadas como dados de entrada na ResNet 3D gerando novos vetores de característica que possuem características não apenas da imagem mas também informações de sua vizinhança. Por fim, os ranqueamentos gerados pelos modelos 2D e 3D foram combinados por meio dos métodos de aprendizado não supervisionado apresentados anteriormente. Válido ressaltar que diversos outros valores de top-n foram experimentados de modo que variamos n de 1 a 15, dado as limitações de dados de entrada da arquitetura utilizada e apresentamos nesse trabalho os resultados mais interessantes obtidos.

5.5.1 Resultados

Para todos os três conjuntos de dados (Stanford 40 *Actions*, Ikizler e Willow *Actions*), as mesmas configurações de parâmetros foram usadas, de modo que os valores dos parâmetros dos algoritmos de *manifold learning* seguiram os valores padrão disponíveis no arcabrouço UDLF (VALEM; PEDRONETTE, 2017).

As Tabelas 4, 5 e 6 apresentam respectivamente os resultados da abordagem proposta nos conjuntos de dados Stanford 40 *Actions*, Ikizler e Willow *Actions*. Podemos observar que a fusão baseada em métodos de ranqueamento manifold (τ_f) apresentou maiores ganhos de eficácia em relação aos modelos isolados (τ_2, τ_3). Também podemos notar que o cenário com tamanho de vizinhança $k_r = 5$ obteve os melhores resultados na maioria dos cenários considerados (por tal os exemplos desta configuração são discutidos na análise visual - Seção 5.5.2). Embora em geral, o uso do método BFS-Tree tenha apresentado melhores resultados do que o LHRR, é possível notar que para os conjuntos de dados Stanford 40 *Actions* e Ikizler na métrica de precisão ($P@x$) o LHRR supera o BFS-Tree para alguns valores de x (para valores de x igual ou inferior a 10 para o Ikizler e inferior ou igual a 15 para o Stanford 40 *Actions*).

Os ganhos relativos obtidos pela abordagem proposta com base na fusão de ranque-

amentos são significativos em todos os cenários. Particularmente, para a métrica MAP, no conjunto de dados Stanford 40 *Actions*, os ganhos absolutos chegam a +12,73% em relação ao modelo 2D (22,36% a 35,09%) e +9,72% em relação ao modelo 3D (25,37% a 35,09%). Considerando o conjunto de dados Ikizler e a métrica MAP, o ganho absoluto é de até +12,61% em relação ao modelo 2D (33,65% a 46,26%) e 1,98% em relação ao modelo 3D (44,28% a 46,26%). Finalmente, no conjunto de dados Willow *Action*, os ganhos absolutos chegam a +14,9% em relação ao modelo 2D (47,47% a 62,37%) e +9,68% em relação ao modelo 3D (52,69% a 62,37%). Considerando os ganhos relativos, os resultados são ainda mais expressivos, com ganhos de até +56,93%, +37,47% e +31,38% nos conjuntos de dados Stanford, Ikizler e Willow, respectivamente.

	Fusion	**P@5**	**P@10**	**P@15**	**P@20**	**P@30**	**P@50**	**P@100**	**MAP**
2D CNN	-	0.6170	0.5422	0.5049	0.4805	0.4476	0.4052	0.3431	0.2236
3D CNN ($k_r = 5$)	-	0.5653	0.4936	0.4626	0.4440	0.4203	0.3911	0.3490	0.2537
3D CNN ($k_r = 10$)	-	0.5431	0.4706	0.4400	0.4218	0.3979	0.3693	0.3283	0.2364
3D CNN ($k_r = 15$)	-	0.5246	0.4503	0.4204	0.4028	0.3797	0.3532	0.3162	0.2297
2D + 3D CNN ($k_r = 5$)	LHRR	**0.6280**	**0.5655**	**0.5374**	**0.5193**	0.4953	0.4635	0.4164	0.3234
2D + 3D CNN ($k_r = 10$)	LHRR	0.6191	0.5557	0.5286	0.5108	0.4850	0.4533	0.4072	0.3129
2D + 3D CNN ($k_r = 15$)	LHRR	0.6051	0.5429	0.5167	0.4982	0.4740	0.4430	0.3983	0.3078
2D + 3D CNN ($k_r = 5$)	BFS-Tree	0.6202	0.5565	0.5306	0.5161	**0.4965**	**0.4730**	**0.4371**	**0.3509**
2D + 3D CNN ($k_r = 10$)	BFS-Tree	0.6088	0.5466	0.5213	0.5062	0.4866	0.4623	0.4275	0.3416
2D + 3D CNN ($k_r = 15$)	BFS-Tree	0.6046	0.5389	0.5115	0.4951	0.4747	0.4522	0.4186	0.3352

Tabela 4 – Resultados da Abordagem De Aprendizado de Representações Sobre o Conjunto de Dados Stanford 40 Actions.

	Fusion	**P@5**	**P@10**	**P@15**	**P@20**	**P@30**	P@50	P@100	MAP
2D CNN	-	0.6578	0.5878	0.5539	0.5331	0.5073	0.4704	0.4136	0.3365
3D CNN ($k_r = 5$)	-	0.6859	0.6272	0.6045	0.5906	0.5698	0.5459	0.5048	0.4428
3D CNN ($k_r = 10$)	-	0.6629	0.6090	0.5871	0.5716	0.5530	0.5271	0.4855	0.4211
3D CNN ($k_r = 15$)	-	0.6258	0.5649	0.5377	0.5235	0.5062	0.4828	0.4437	0.3856
2D + 3D CNN ($k_r = 5$)	LHRR	**0.6902**	**0.6334**	0.6094	0.5947	0.5722	0.5413	0.4962	0.4445
2D + 3D CNN ($k_r = 10$)	LHRR	0.6835	0.6240	0.5980	0.5823	0.5597	0.5297	0.4835	0.4310
2D + 3D CNN ($k_r = 15$)	LHRR	0.6534	0.5982	0.5688	0.5545	0.5362	0.5092	0.4643	0.4115
2D + 3D CNN ($k_r = 5$)	BFS-Tree	0.6798	0.6328	**0.6111**	**0.5961**	**0.5784**	**0.5546**	**0.5198**	**0.4626**
2D + 3D CNN ($k_r = 10$)	BFS-Tree	0.6802	0.6222	0.5997	0.5877	0.5698	0.5454	0.5083	0.4519
2D + 3D CNN ($k_r = 15$)	BFS-Tree	0.6581	0.5978	0.5744	0.5608	0.5422	0.5185	0.4836	0.4275

Tabela 5 – Resultados da Abordagem De Aprendizado de Representações Sobre o Conjunto de Dados Ikizler.

5.5.2 Análise Visual

A fim de enriquecer a discussão sobre a abordagem proposta, empregamos métodos de redução de dimensionalidade para representar o impacto do método em uma projeção 2-D do espaço de *features*. A análise foi realizada nos três conjuntos de dados acima mencionados, utilizando o algoritmo t-SNE (MAATEN; HINTON, 2008) e o algoritmo UMAP (MCINNES et al., 2018).

A Figura 19, mostra as visualizações da aplicação do t-SNE nos conjuntos de dados Stanford-40, Ikizler e Willow Actions. Para cada conjunto de dados, apresenta-se,

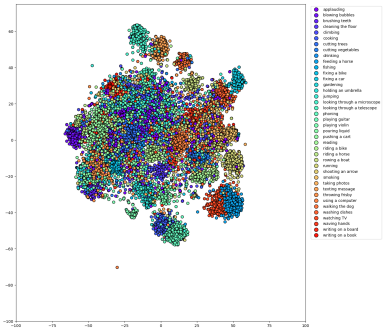
	Fusion	**P@5**	**P@10**	**P@15**	**P@20**	**P@30**	**P@50**	**P@100**	**MAP**
2D CNN	-	0.7745	0.7259	0.6959	0.6776	0.6479	0.5964	0.5011	0.4747
3D CNN ($k_r = 5$)	-	0.7524	0.7069	0.6859	0.6712	0.6483	0.6172	0.5463	0.5269
3D CNN ($k_r = 10$)	-	0.7324	0.6835	0.6591	0.6444	0.6207	0.5877	0.5197	0.5035
3D CNN ($k_r = 15$)	-	0.7183	0.6748	0.6553	0.6435	0.6261	0.6048	0.5434	0.5254
2D + 3D CNN ($k_r = 5$)	LHRR	0.7842	0.7423	0.7240	0.7106	0.6915	0.6628	0.5942	0.5924
2D + 3D CNN ($k_r = 10$)	LHRR	0.7748	0.7325	0.7172	0.7003	0.6785	0.6473	0.5786	0.5775
2D + 3D CNN ($k_r = 15$)	LHRR	0.7701	0.7268	0.7096	0.7003	0.6774	0.6434	0.5846	0.5801
2D + 3D CNN ($k_r = 5$)	BFS-Tree	**0.7897**	**0.7447**	**0.7301**	**0.7168**	**0.7037**	**0.6773**	**0.6229**	**0.6237**
2D + 3D CNN ($k_r = 10$)	BFS-Tree	0.7719	0.7360	0.7198	0.7059	0.6873	0.6589	0.6044	0.6048
2D + 3D CNN ($k_r = 15$)	BFS-Tree	0.7706	0.7338	0.7169	0.7058	0.6878	0.6588	0.6109	0.6086

Tabela 6 – Resultados da Abordagem De Aprendizado de Representações Sobre o Conjunto de Dados Willow.

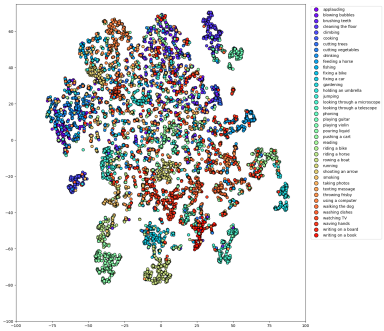
respectivamente, a distância obtida a partir das características do modelo 2D, seguida da fusão dos rankings do modelo 2D e 3D pelos algoritmos LHRR e BFS-TRee. Como podemos notar nas representações, ambas as abordagens baseadas na fusão de classificação múltipla resultaram em melhor separabilidade das classes para todos os cenários.

Por sua vez a Figura 20 mostra as visualizações do método UMAP para os conjuntos de dados. Novamente, para cada conjunto nós temos respectivamente o cenário baseado nas características do modelo 2D, seguidas da fusão de ambos os modelos utilizando do LHRR, seguida da fusão utilizando o método BFS-Tree. De mesmo modo que notado para o t-SNE, é possível notar uma melhor separabilidade das classes para ambos os métodos de aprendizado não supervisionados sobre todos os conjuntos de dados, significando que a nova distância é uma medida mais representativa das características do conjunto de dados e consequentemente mais eficaz para recuperação.

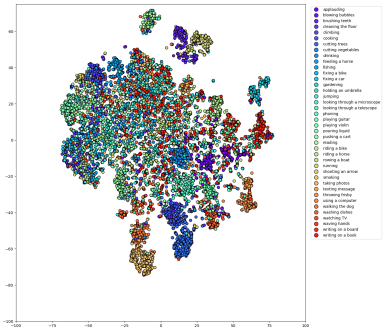
Em outra análise visual para avaliar a eficácia da abordagem proposta, as Figuras 21, 22, e 23 ilustram o listas ranqueadas calculadas pelo modelo 2D CNN e pela fusão dos ranqueamentos dos modelos 2D e 3D pelo método BFS-Tree. As bordas vermelhas indicam imagens que não pertencem à mesma classe da imagem de consulta. Neste conjunto de representações, é possível visualizar que o impacto de nossa abordagem é especialmente notável para certas instâncias de cada conjunto de dados.



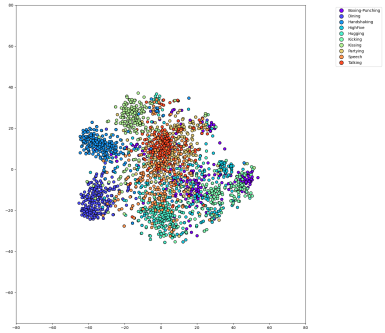
Stanford ResNet 2D.



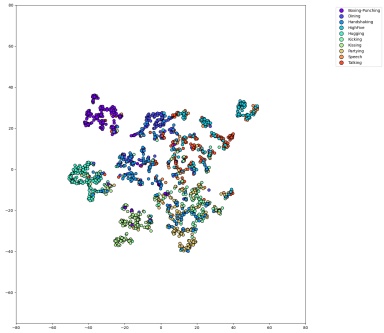
Stanford ResNet 2D+3D (LHRR).



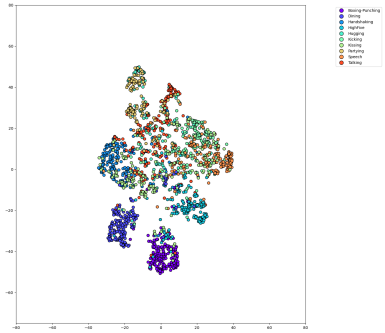
Stanford ResNet2D+3D (BFS-Tree).



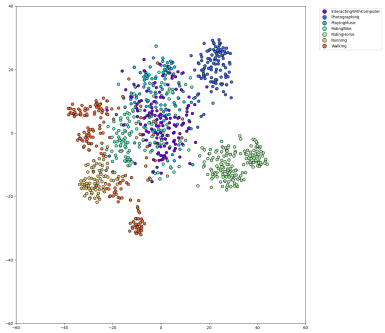
Ikizler ResNet 2D.



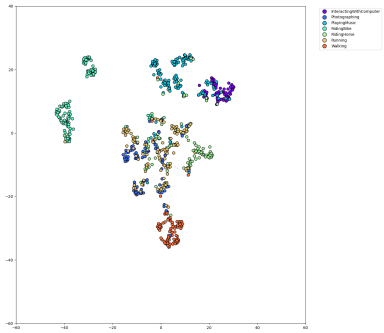
Ikizler ResNet 2D+3D (LHRR).



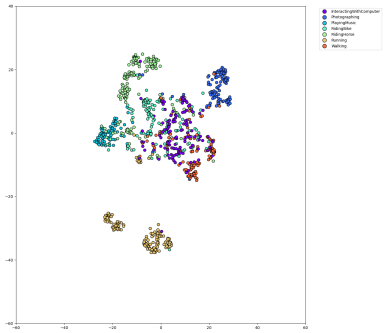
Ikizler ResNet 2D+3D (BFS-Tree).



Willow ResNet 2D.

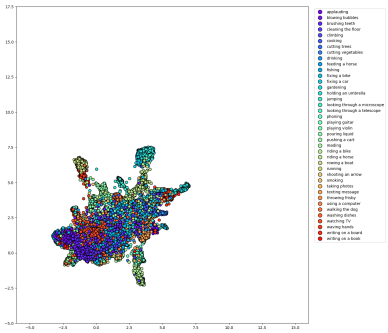


Willow ResNet 2D+3D (LHRR).

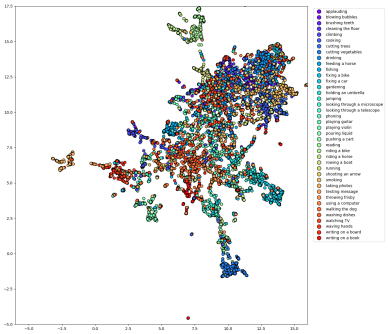


Willow ResNet 2D+3D (BFS-Tree).

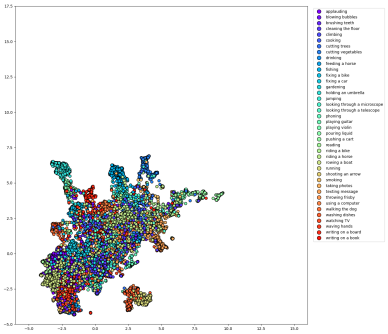
Figura 19 – Visualização t-SNE da Abordagem de Aprendizado de Representações em Comparação Com o Modelo Original.



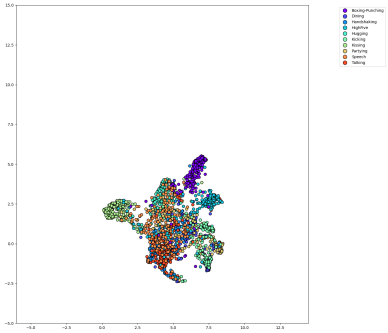
Stanford ResNet 2D.



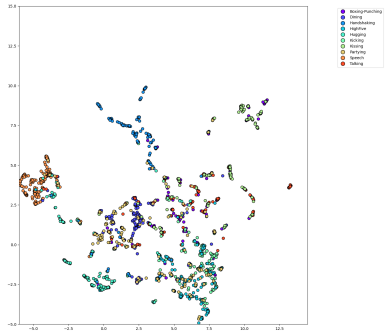
Stanford ResNet 2D+3D (LHRR).



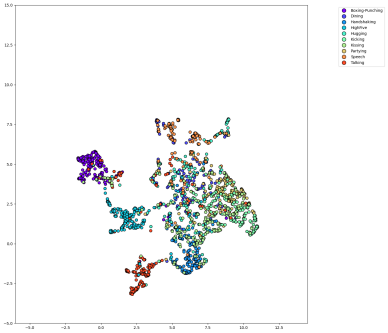
Stanford ResNet 2D+3D (BFS-Tree).



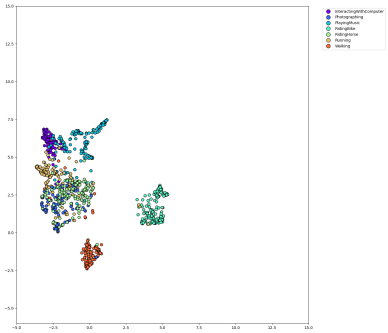
Ikizler ResNet 2D.



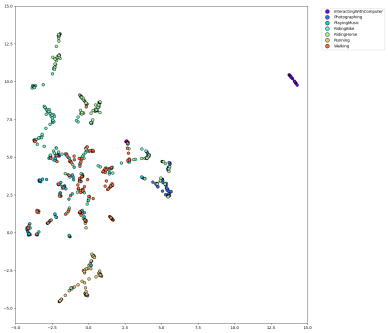
Ikizler ResNet 2D+3D (LHRR).



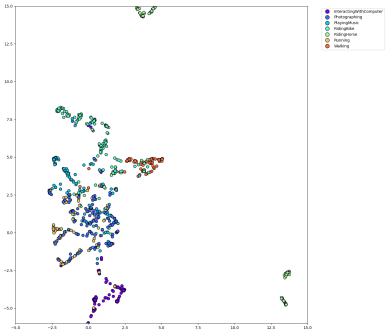
Ikizler ResNet 2D+3D (BFS-Tree).



Willow ResNet 2D.



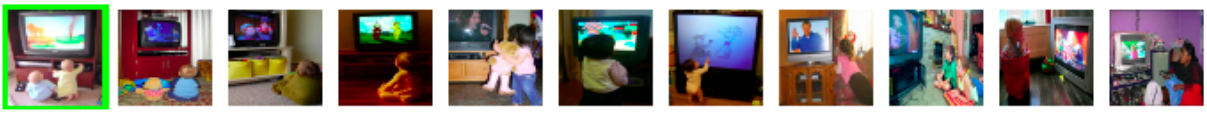
Willow ResNet 2D+3D (LHRR).



Willow ResNet 2D+3D (BFS-Tree).

Figura 20 – Visualização UMAP da Abordagem de Aprendizado de Representações em Comparação Com o Modelo Original.

Resultados da Recuperação pelos vetores de características da ResNet 2D.



Resultados da Recuperação Utilizando as Informações da ResNet 2D e 3D Combinadas pelo BFS-Tree.

Figura 21 – Resultados Visuais da Recuperação no conjunto de dados Stanford 40 Actions, Utilizando a Representação 2D e a Abordagem Proposta. Imagem de Consulta em Bordas Verdes, Imagens que Não Pertencem a Mesma Classe em Bordas Vermelhas.

Resultados da Recuperação pelos vetores de características da ResNet 2D.

Resultados da Recuperação Utilizando as Informações da ResNet 2D e 3D Combinadas pelo BFS-Tree.

Figura 22 – Resultados Visuais da Recuperação no conjunto de dados Ikizler, Utilizando a Representação 2D e a Abordagem Proposta. Imagem de Consulta em Bordas Verdes, Imagens que Não Pertencem a Mesma Classe em Bordas Vermelhas.

Resultados da Recuperação pelos vetores de características da ResNet 2D.

Resultados da Recuperação Utilizando as Informações da ResNet 2D e 3D Combinadas pelo BFS-Tree.

Figura 23 – Resultados Visuais da Recuperação no conjunto de dados Willow Actions, Utilizando a Representação 2D e a Abordagem Proposta. Imagem de Consulta em Bordas Verdes, Imagens que Não Pertencem a Mesma Classe em Bordas Vermelhas.

6 Conclusões

Considerando o crescimento vertiginoso de coleções multimídia e a escassez ou ausência de dados rotulados, a necessidade de investigação de métodos de recuperação baseada no conteúdo capazes de operar nestes cenários torna-se premente. Além disso, a possibilidade de explorar informações complementares de diferentes modalidades representa um potencial promissor. Esta dissertação apresentou como contribuições centrais três abordagens para recuperação multimodal de vídeos e uma abordagem de recuperação com representações multimodais de imagens. Além da multimodalidade, as abordagens propostas tem foco em aprendizado de representações e aspectos de similaridade contextual.

Dentre as abordagens de recuperação de vídeo, uma delas utiliza somente métodos de *manifold learning*, enquanto as outras duas combinam métodos de *manifold learning* e Redes Convolucionais baseadas em Grafos. As redes são treinadas de modo totalmente não-supervisionado utilizando o algoritmo *Deep Graph Infomax* para recuperação multimodal. As três abordagens de recuperação de vídeo apresentaram resultados promissores. De modo geral, as abordagens que utilizaram GCNs obtiveram resultados superiores na maioria dos cenários. As abordagens de recuperação de vídeo propostas derivaram um artigo submetido e em fase de avaliação para a conferência *IEEE International Conference in Image Processing - ICIP 2022* (ALMEIDA; VALEM; PEDRONETTE, 2022).

A abordagem para recuperação de imagens baseia-se em aprendizado de representações e métodos *manifold learning*. A abordagem proposta utiliza uma CNN 3D pré-treinada para criação de representações que possuem informações de vizinhança para aprimorar a representação de cada imagem. Os resultados também apresentaram ganhos significativos para os três conjuntos de dados considerados. Tal abordagem também originou um artigo científico publicado no *SIBGRAPI – Conference on Graphics, Patterns and Images 2021* (ALMEIDA et al., 2021).

Como trabalhos futuros podemos mencionar: a avaliação de outros modelos baseados em grafos para a tarefa de criação de *embenddings*; o uso de diferentes descritores para criação dos *embenddings* iniciais, tal como EfficientNets (TAN; LE, 2019) ; a análise de diferentes métricas de eficácia de recuperação; a avaliação diferentes métodos de agregação. Pretende-se investigar também abordagens para a criação de *embenddings* em cenários semi-supervisionados e auto-supervisionados. Em cenários semi-supervisionados, pretende-se investigar o treinamento da rede utilizando apenas os elementos mais representativos de cada classe. Para o cenário auto-supervisionado, cogita-se criar pseudo-rótulos a partir da distância dos elementos e dados de agrupamentos.

Além disso, seria interessante incorporar a modalidade de texto ao processo multi-

modal, seja a partir de legendas presentes em vídeos ou anotações disponibilizadas junto com as coleções de dados ou realizadas automaticamente com o uso de aprendizado de máquina. Para as abordagens de vídeo futuramente seria relevante utilizar de mais de um quadro para extração de características da modalidade de imagem, podendo ser utilizados alguns quadros chave para tal, para que se tenha uma representação que melhor sumarize o vídeo.

Por fim, considerando trabalhos futuros para cenários multimodais seria expressivo explorar a representação de um espaço comum de representação, criando com isso relações entre as diferentes modalidades para cada instância em apenas um espaço inicial de *features* e realizando o processo de recuperação sobre ele.

Referências

ABADAL, S. et al. Computing graph neural networks: A survey from algorithms to accelerators. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 54, n. 9, oct 2021. ISSN 0360-0300. Nenhuma citação no texto.

ADCOCK, J. et al. Fxpal experiments for trecvid 2004. 01 2004. Nenhuma citação no texto.

AGARWAL, M.; MOSTAFA, J. Content-based image retrieval for alzheimer's disease detection. In: *2011 9th International Workshop on Content-Based Multimedia Indexing (CBMI)*. [S.l.: s.n.], 2011. p. 13–18. ISSN 1949-3983. Nenhuma citação no texto.

AH-PINE, J.; CSURKA, G.; CLINCHANT, S. Unsupervised visual and textual information fusion in cbmir using graph-based methods. *ACM Transactions on Information Systems*, v. 33, p. 1–31, 02 2015. Nenhuma citação no texto.

Almeida, J.; Leite, N. J.; da S. Torres, R. Comparison of video sequences with histograms of motion patterns. In: *2011 18th IEEE International Conference on Image Processing*. [S.l.: s.n.], 2011. p. 3673–3676. Nenhuma citação no texto.

ALMEIDA, J.; PEDRONETTE, D. C. G.; PENATTI, O. A. B. Unsupervised manifold learning for video genre retrieval. In: BAYRO-CORROCHANO, E.; HANCOCK, E. (Ed.). *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Cham: Springer International Publishing, 2014. p. 604–612. ISBN 978-3-319-12568-8. Nenhuma citação no texto.

ALMEIDA, J.; VALEM, L.; PEDRONETTE, D. A rank aggregation framework for video interestingness prediction. In: . [S.l.: s.n.], 2017. p. 3–14. ISBN 978-3-319-68559-5. Nenhuma citação no texto.

ALMEIDA, L. B. de et al. Representation learning for image retrieval through 3d cnn and manifold ranking. In: *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. [S.l.: s.n.], 2021. p. 417–424. Nenhuma citação no texto.

ALMEIDA, L. B. de; VALEM, L. P.; PEDRONETTE, D. C. G. Graph convolutional networks and manifold ranking for multimodal video retrieval. In: *International Conference in Image Processing*. [S.l.: s.n.], 2022. Nenhuma citação no texto.

AMIR, A. et al. Ibm research trecvid-2003 video retrieval system. In: *In NIST TRECVID-2003*. [S.l.: s.n.], 2003. Nenhuma citação no texto.

ATREY, P. K. et al. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, v. 16, n. 6, p. 345–379, Nov 2010. ISSN 1432-1882. Disponível em: <https://doi.org/10.1007/s00530-010-0182-0>. Nenhuma citação no texto.

ATWOOD, J.; TOWSLEY, D. Search-convolutional neural networks. *CoRR*, abs/1511.02136, 2015. Disponível em: <http://arxiv.org/abs/1511.02136>. Nenhuma citação no texto.

BABENKO, A. et al. Neural codes for image retrieval. In: _____. *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*. Cham: Springer International Publishing, 2014. p. 584–599. ISBN 978-3-319-10590-1. Disponível em: <https://doi.org/10.1007/978-3-319-10590-1_38>. Nenhuma citação no texto.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Recuperação de Informação - 2ed: Conceitos e Tecnologia das Máquinas de Busca*. Bookman Editora, 2013. ISBN 9788582600498. Disponível em: <https://books.google.com.br/books?id=YWk3AgAAQBAJ>. Nenhuma citação no texto.

BAJI, F.; MOCANU, M. Chain code approach for shape based image retrieval. *Indian Journal of Science and Technology*, v. 11, p. 1–17, 01 2018. Nenhuma citação no texto.

Bashir, F. I.; Khokhar, A. A.; Schonfeld, D. Real-time motion trajectory-based indexing and retrieval of video sequences. *IEEE Transactions on Multimedia*, v. 9, n. 1, p. 58–65, 2007. Nenhuma citação no texto.

BAY, H. et al. Speeded-up robust features (surf). *Computer vision and image understanding*, Elsevier, v. 110, n. 3, p. 346–359, 2008. Nenhuma citação no texto.

Benam, A.; Drew, M. S.; Atkins, M. S. A cbir system for locating and retrieving pigment network in dermoscopy images using dermoscopy interest point detection. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. [S.l.: s.n.], 2017. p. 122–125. Nenhuma citação no texto.

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 35, n. 8, p. 1798–1828, 2013. Nenhuma citação no texto.

BHOWMIK, N. et al. Efficient fusion of multidimensional descriptors for image retrieval. In: *2014 IEEE International Conference on Image Processing (ICIP)*. [S.l.: s.n.], 2014. p. 5766–5770. ISSN 1522-4880. Nenhuma citação no texto.

Boureau, Y. et al. Learning mid-level features for recognition. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2010. p. 2559–2566. Nenhuma citação no texto.

CALONDER, M. et al. Brief: Binary robust independent elementary features. In: _____. *Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 778–792. ISBN 978-3-642-15561-1. Disponível em: <https://doi.org/10.1007/978-3-642-15561-1_56>. Nenhuma citação no texto.

CAMPOS, V.; PEDRONETTE, D. Effective speaker retrieval and recognition through vector quantization and unsupervised distance learning. In: . [S.l.: s.n.], 2016. p. 27–32. Nenhuma citação no texto.

Carlos Guimarães Pedronette, D.; VALEM, L. P.; TORRES, R. da S. A bfs-tree of ranking references for unsupervised manifold learning. *Pattern Recognition*, v. 111, p. 107666, 2021. Nenhuma citação no texto.

CHEN, H. et al. Vggsound: A large-scale audio-visual dataset. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2020. p. 721–725. Nenhuma citação no texto.

CHUNG, Y.; WENG, W. Learning deep representations of medical images using siamese cnns with application to content-based image retrieval. *CoRR*, abs/1711.08490, 2017. Disponível em: <http://arxiv.org/abs/1711.08490>. Nenhuma citação no texto.

COOKE, E. et al. Trecvid 2004 experiments in dublin city university. In: *In NIST TRECVID*. [S.l.: s.n.], 2004. Nenhuma citação no texto.

CORMACK, G. V.; CLARKE, C. L. A.; BUETTCHER, S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2009. (SIGIR '09), p. 758–759. ISBN 978-1-60558-483-6. Disponível em: <http://doi.acm.org/10.1145/1571941.1572114>. Nenhuma citação no texto.

DATTA, R. et al. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, ACM, New York, NY, USA, v. 40, n. 2, p. 5:1–5:60, 2008. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/1348246.1348248>. Nenhuma citação no texto.

Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. 28, n. 4, p. 357–366, 1980. Nenhuma citação no texto.

DELAITRE, V.; LAPTEV, I.; SIVIC, J. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: . [S.l.: s.n.], 2010. Updated version, available at http://www.di.ens.fr/willow/research/stillactions/. Nenhuma citação no texto.

DENG, J. et al. ImageNet: A Large-Scale Hierarchical Image Database. In: *CVPR09*. [S.l.: s.n.], 2009. Nenhuma citação no texto.

DENG, L. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, Cambridge University Press, v. 3, 2014. Nenhuma citação no texto.

DOURADO, I. C.; PEDRONETTE, D. C. G.; TORRES, R. da S. Unsupervised graph-based rank aggregation for improved retrieval. *Information Processing Management*, v. 56, n. 4, p. 1260 – 1279, 2019. ISSN 0306-4573. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0306457318307647>. Nenhuma citação no texto.

DOURADO, I. C.; TABBONE, S.; TORRES, R. Multimodal representation model based on graph-based rank fusion. *ArXiv*, abs/1912.10314, 2019. Nenhuma citação no texto.

DOURADO, I. C.; TABBONE, S.; TORRES, R. da S. *Multimodal Prediction based on Graph Representations*. 2020. Nenhuma citação no texto.

Fablet, R.; Bouthemy, P.; Perez, P. Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval. *IEEE Transactions on Image Processing*, v. 11, n. 4, p. 393–407, 2002. Nenhuma citação no texto.

FOLEY, C. et al. Trecvid 2005 experiments at dublin city university. In: OVER, P. et al. (Ed.). *2005 TREC Video Retrieval Evaluation, TRECVID 2005, Gaithersburg, MD, USA, November 14-15, 2005*. National Institute of Standards and Technology (NIST), 2005. Disponível em: <https://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/dcu.pdf>. Nenhuma citação no texto.

FOOTE, J. An overview of audio information retrieval. *Multimedia Systems*, v. 7, n. 1, p. 2–10, Jan 1999. ISSN 1432-1882. Disponível em: <https://doi.org/10.1007/s005300050106>. Nenhuma citação no texto.

FORCéN, J. et al. *Co-occurrence of deep convolutional features for image search*. 2020. Nenhuma citação no texto.

Fu, Z. et al. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, v. 13, n. 2, p. 303–319, 2011. Nenhuma citação no texto.

GALLICCHIO, C.; MICHELI, A. Graph echo state networks. In: . [S.l.: s.n.], 2010. p. 1 – 8. Nenhuma citação no texto.

GAMMULLE, H. et al. Two stream lstm : A deep fusion framework for human action recognition. In: . [S.l.: s.n.], 2017. Nenhuma citação no texto.

GEMMEKE, J. F. et al. Audio set: An ontology and human-labeled dataset for audio events. In: *Proc. IEEE ICASSP 2017*. New Orleans, LA: [s.n.], 2017. Nenhuma citação no texto.

GILMER, J. et al. Neural message passing for quantum chemistry. *CoRR*, abs/1704.01212, 2017. Disponível em: <http://arxiv.org/abs/1704.01212>. Nenhuma citação no texto.

GOLDMAN-EISLER, F. Speech analysis and mental processes. *Language and Speech*, v. 1, n. 1, p. 59–75, 1958. Disponível em: <https://doi.org/10.1177/002383095800100105>. Nenhuma citação no texto.

GORI, M.; MONFARDINI, G.; SCARSELLI, F. A new model for learning in graph domains. In: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* [S.l.: s.n.], 2005. v. 2, p. 729–734 vol. 2. Nenhuma citação no texto.

GU, Y. et al. Deep graph-based multimodal feature embedding for endomicroscopy image retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, v. 32, n. 2, p. 481–492, 2021. Nenhuma citação no texto.

GUO, K. et al. An effective and economical architecture for semantic-based heterogeneous multimedia big data retrieval. *Journal of Systems and Software*, v. 102, p. 207 – 216, 2015. ISSN 0164-1212. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0164121214002040>. Nenhuma citação no texto.

HAMMOND, D. K.; VANDERGHEYNST, P.; GRIBONVAL, R. *Wavelets on Graphs via Spectral Graph Theory*. 2009. Nenhuma citação no texto.

Han, X. et al. Vrfp: On-the-fly video retrieval using web images and fast fisher vector products. *IEEE Transactions on Multimedia*, v. 19, n. 7, p. 1583–1595, 2017. Nenhuma citação no texto.

HAUPTMANN, A. et al. Informedia at trecvid 2003: Analyzing and searching broadcast news video. 01 2003. Nenhuma citação no texto.

HAYAKAWA, Y. et al. Feature extraction of video using deep neural network. In: . [S.l.: s.n.], 2016. p. 465–470. Nenhuma citação no texto.

HE, K. et al. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. Disponível em: <http://arxiv.org/abs/1512.03385>. Nenhuma citação no texto.

HE, K. et al. Deep residual learning for image recognition. In: *IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR'16)*. [S.l.: s.n.], 2016. p. 770–778. Nenhuma citação no texto.

HJELM, R. D. et al. *Learning deep representations by mutual information estimation and maximization*. 2019. Nenhuma citação no texto.

HOI, S. C.; LIU, W.; CHANG, S.-F. Semi-supervised distance metric learning for collaborative image retrieval and clustering. *ACM Transactions on Multimedia Computing and Communication Applications*, v. 6, n. 3, p. 18:1–18:26, August 2010. ISSN 1551-6857. Nenhuma citação no texto.

HOU, R.; CHEN, C.; SHAH, M. Tube convolutional neural network (t-cnn) for action detection in videos. 03 2017. Nenhuma citação no texto.

HOWARD, C. G. Speech analysis-synthesis scheme using continuous parameters. *The Journal of the Acoustical Society of America*, v. 28, n. 6, p. 1091–1098, 1956. Disponível em: <https://doi.org/10.1121/1.1908565>. Nenhuma citação no texto.

HU, W. et al. A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, v. 41, p. 797–819, 11 2011. Nenhuma citação no texto.

HUANG, G.; LIU, Z.; WEINBERGER, K. Q. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. Nenhuma citação no texto.

HUANG, J. et al. Image indexing using color correlograms. In: *IEEE Int. Conf. Computer Vision and Pattern Recognition (CVPR'97)*. [S.l.: s.n.], 1997. p. 762–768. Nenhuma citação no texto.

Hui, T.; Tang, X.; Loy, C. C. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 8981–8989. Nenhuma citação no texto.

Ilg, E. et al. Flownet 2.0: Evolution of optical flow estimation with deep networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2017. p. 1647–1655. Nenhuma citação no texto.

IONESCU, B. et al. *Fusion in Computer Vision: Understanding Complex Visual Content*. [S.l.]: Springer Publishing Company, Incorporated, 2014. ISBN 3319056956, 9783319056951. Nenhuma citação no texto.

JI, S. et al. 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, v. 35, n. 1, p. 221–231, 2013. Nenhuma citação no texto.

JUNG, Y.-K.; LEE, K.-W.; HO, Y.-S. Content-based event retrieval using semantic scene interpretation for automated traffic surveillance. *Intelligent Transportation Systems, IEEE Transactions on*, v. 2, p. 151 – 163, 10 2001. Nenhuma citação no texto.

KAUR, M. A.; KAUR, M. R. Feature extraction from video data for indexing and retrieval. In: . [S.l.: s.n.], 2015. Nenhuma citação no texto.

KAY, W. et al. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. Disponível em: <http://arxiv.org/abs/1705.06950>. Nenhuma citação no texto.

KIPF, T. N.; WELLING, M. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016. Disponível em: <http://arxiv.org/abs/1609.02907>. Nenhuma citação no texto.

KOFLER, C.; LARSON, M.; HANJALIC, A. User intent in multimedia search: A survey of the state of the art and future challenges. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 49, n. 2, p. 36:1–36:37, ago. 2016. ISSN 0360-0300. Disponível em: <http://doi.acm.org/10.1145/2954930>. Nenhuma citação no texto.

KONG, Q. et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, v. 28, p. 2880–2894, 2020. Nenhuma citação no texto.

KOVALEV, V.; VOLMER, S. Color co-occurence descriptors for querying-by-example. In: *ICMM*. [S.l.: s.n.], 1998. p. 32. ISBN 0-8186-8911-0. Nenhuma citação no texto.

Krueger, A.; Haeb-Umbach, R. Model-based feature enhancement for reverberant speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, v. 18, n. 7, p. 1692–1707, 2010. Nenhuma citação no texto.

LI, L. et al. A rank aggregation framework for video multimodal geocoding. *Multimedia Tools and Applications*, v. 73, p. 1323–1359, 12 2014. Nenhuma citação no texto.

LI, S.; DENG, W. Deep facial expression recognition: A survey. 04 2018. Nenhuma citação no texto.

LI, T.; OGIHARA, M.; LI, Q. A comparative study on content-based music genre classification. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. New York, NY, USA: Association for Computing Machinery, 2003. (SIGIR '03), p. 282–289. ISBN 1581136463. Disponível em: <https://doi.org/10.1145/860435.860487>. Nenhuma citação no texto.

LI YANXIONG, Z. X. et al. Using multi-stream hierarchical deep neural network to extract deep audio feature for acoustic event detection. *Multimedia Tools and Applications*, v. 77, 01 2018. Nenhuma citação no texto.

LING, H.; YANG, X.; LATECKI, L. J. Balancing deformability and discriminability for shape matching. In: *ECCV*. [S.l.: s.n.], 2010. v. 3, p. 411–424. Nenhuma citação no texto.

LOWE, D. Object recognition from local scale-invariant features. In: *ICCV*. [S.l.: s.n.], 1999. p. 1150–1157. Nenhuma citação no texto.

MAATEN, L. van der; HINTON, G. Viualizing data using t-sne. *Journal of Machine Learning Research*, v. 9, p. 2579–2605, 11 2008. Nenhuma citação no texto.

MCINNES, L. et al. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, v. 3, n. 29, p. 861, 2018. Nenhuma citação no texto.

MERLIN, V. The axiomatic characterization of majority voting and scoring rules. *Mathématiques et sciences humaines*, 02 2003. Nenhuma citação no texto.

MICHELI, A. Neural network for graphs: A contextual constructive approach. *IEEE Transactions on Neural Networks*, v. 20, n. 3, p. 498–511, 2009. Nenhuma citação no texto.

MISRAA, A. K. et al. *Multi-Modal Retrieval using Graph Neural Networks*. 2020. Nenhuma citação no texto.

MITROVIC, D.; ZEPPELZAUER, M.; BREITENEDER, C. Discrimination and retrieval of animal sounds. In: . [S.l.: s.n.], 2006. p. 5 pp. Nenhuma citação no texto.

MITROVIC, D.; ZEPPELZAUER, M.; BREITENEDER, C. Features for content-based audio retrieval. *Advances in Computers*, v. 78, p. 71–150, 01 2010. Nenhuma citação no texto.

MONFORT, M. et al. Moments in time dataset: one million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–8, 2019. ISSN 0162-8828. Nenhuma citação no texto.

MONFORT, M. et al. *Multi-Moments in Time: Learning and Interpreting Models for Multi-Action Video Understanding*. 2019. Nenhuma citação no texto.

MOURãO, A.; MARTINS, F.; MAGALHãES, J. Multimodal medical information retrieval with unsupervised rank fusion. *Computerized Medical Imaging and Graphics*, 05 2014. Nenhuma citação no texto.

NIEPERT, M.; AHMED, M.; KUTZKOV, K. Learning convolutional neural networks for graphs. *CoRR*, abs/1605.05273, 2016. Disponível em: <http://arxiv.org/abs/1605.05273>. Nenhuma citação no texto.

NING, H.; ZHAO, B.; YUAN, Y. *Semantics-Consistent Representation Learning for Remote Sensing Image-Voice Retrieval*. 2021. Nenhuma citação no texto.

OJALA, T.; PIETIKäINEN, M.; MäENPää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI*, v. 24, n. 7, p. 971–987, 2002. ISSN 0162-8828. Nenhuma citação no texto.

Panyapanuwat, P.; Kamonsantiroj, S.; Pipanmaekaporn, L. Unsupervised learning hash for content-based audio retrieval using deep neural networks. In: *2019 11th International Conference on Knowledge and Smart Technology (KST)*. [S.l.: s.n.], 2019. p. 99–104. Nenhuma citação no texto.

PATRON-PEREZ, A. et al. Structured learning of human interactions in tv shows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 34, n. 12, p. 2441–2453, 2012. Nenhuma citação no texto.

PEDRONETTE, D. C. G.; GONçALVES, F. M. F.; GUILHERME, I. R. Unsupervised manifold learning through reciprocal knn graph and connected components for image retrieval tasks. *Pattern Recognition*, v. 75, p. 161 – 174, 2018. Nenhuma citação no texto.

PEDRONETTE, D. C. G.; GONçALVES, F. M. F.; GUILHERME, I. R. Unsupervised manifold learning through reciprocal knn graph and connected components for image retrieval tasks. *Pattern Recognition*, v. 75, p. 161 – 174, 2018. ISSN 0031-3203. Distance Metric Learning for Pattern Recognition. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320317301978>. Nenhuma citação no texto.

PEDRONETTE, D. C. G.; GONçALVES, F. M. F.; GUILHERME, I. R. Unsupervised manifold learning through reciprocal knn graph and connected components for image retrieval tasks. *Pattern Recognition*, v. 75, p. 161–174, 2018. ISSN 0031-3203. Distance Metric Learning for Pattern Recognition. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320317301978>. Nenhuma citação no texto.

PEDRONETTE, D. C. G.; TORRES, R. da S. Shape retrieval using contour features and distance optmization. In: *VISAPP*. Angers, France: [s.n.], 2010. v. 1, p. 197 – 202. Nenhuma citação no texto.

PEDRONETTE, D. C. G.; TORRES, R. da S. Unsupervised rank diffusion for content-based image retrieval. *Neurocomputing*, v. 260, p. 478 – 489, 2017. ISSN 0925-2312. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231217308330>. Nenhuma citação no texto.

PEDRONETTE, D. C. G. et al. Multimedia retrieval through unsupervised hypergraph-based manifold ranking. *IEEE Trans. Image Processing*, IEEE, v. 28, n. 12, p. 5824–5838, 2019. Nenhuma citação no texto.

PELTONEN, V. et al. Computational auditory scene recognition. *Proceedings of the IEEE*, v. 2, 06 2002. Nenhuma citação no texto.

PENATTI, O. A.; VALLE, E.; TORRES, R. da S. Comparative study of global color and texture descriptors for web image retrieval. *Journal of Visual Communication and Image Representation*, v. 23, n. 2, p. 359 – 380, 2012. ISSN 1047-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1047320311001465>. Nenhuma citação no texto.

PENG, Y.; ZHAO, Y.; ZHANG, J. Two-stream collaborative learning with spatial-temporal attention for video classification. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 29, 11 2017. Nenhuma citação no texto.

PIRAS, L.; GIACINTO, G. Information fusion in content based image retrieval: A comprehensive overview. *Information Fusion*, v. 37, 01 2017. Nenhuma citação no texto.

PIRAS, L.; TRONCI, R.; GIACINTO, G. Diversity in ensembles of codebooks for visual concept detection. In: _____. *Image Analysis and Processing – ICIAP 2013: 17th International Conference, Naples, Italy, September 9-13, 2013, Proceedings, Part II*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 399–408. ISBN 978-3-642-41184-7. Disponível em: <https://doi.org/10.1007/978-3-642-41184-7_41>. Nenhuma citação no texto.

RIAN, Z.; CHRISTANTI, V.; HENDRYLI, J. Content-based image retrieval using convolutional neural networks. In: *2019 IEEE International Conference on Signals and Systems (ICSigSys)*. [S.l.: s.n.], 2019. p. 1–7. Nenhuma citação no texto.

Riccardi, G.; Hakkani-Tur, D. Active learning: theory and applications to automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, v. 13, n. 4, p. 504–511, 2005. Nenhuma citação no texto.

RUBLEE, E. et al. Orb: An efficient alternative to sift or surf. In: *2011 International Conference on Computer Vision*. [S.l.: s.n.], 2011. p. 2564–2571. ISSN 1550-5499. Nenhuma citação no texto.

SALAU, A. O.; JAIN, S. Feature extraction: A survey of the types, techniques, applications. In: *2019 International Conference on Signal Processing and Communication (ICSC)*. [S.l.: s.n.], 2019. p. 158–164. Nenhuma citação no texto.

SANTOS, S. F.; ALMEIDA, J. Faster and accurate compressed video action recognition straight from the frequency domain. In: *Conference on Graphics, Patterns and Images (SIBGRAPI'20)*. [S.l.: s.n.], 2020. p. 62–68. Nenhuma citação no texto.

SANTOS, S. F.; SEBE, N.; ALMEIDA, J. CV-C3D: action recognition on compressed videos with convolutional 3d networks. In: *Conference on Graphics, Patterns and Images (SIBGRAPI'19)*. [S.l.: s.n.], 2019. p. 24–30. Nenhuma citação no texto.

SCARSELLI, F. et al. The graph neural network model. *Trans. Neur. Netw.*, IEEE Press, v. 20, n. 1, p. 61–80, jan 2009. ISSN 1045-9227. Disponível em: <https://doi.org/10.1109/TNN.2008.2005605>. Nenhuma citação no texto.

SCHOBER, J. pierre; HERMES, T.; HERZOG, O. Content-based image retrieval by ontology-based object recognition. In: *in KI-2004 Workshop on Applications of Description Logics*. [S.l.: s.n.], 2004. p. 61–67. Nenhuma citação no texto.

SEJDIC, E.; DJUROVIC, I.; JIANG, J. Time–frequency feature representation using energy concentration: An overview of recent advances. *Digital Signal Processing*, v. 19, p. 153–183, 01 2009. Nenhuma citação no texto.

SERMANET, P.; CHINTALA, S.; LECUN, Y. Convolutional neural networks applied to house numbers digit classification. *CoRR*, abs/1204.3968, 2012. Disponível em: <http://arxiv.org/abs/1204.3968>. Nenhuma citação no texto.

SHARMA, G.; UMAPATHY, K.; KRISHNAN, S. Trends in audio signal feature extraction methods. *Applied Acoustics*, v. 158, p. 107020, 2020. ISSN 0003-682X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0003682X19308795>. Nenhuma citação no texto.

SHEN, Z. et al. Semi-supervised graph convolutional hashing network for large-scale cross-modal retrieval. In: *2020 IEEE International Conference on Image Processing (ICIP)*. [S.l.: s.n.], 2020. p. 2366–2370. Nenhuma citação no texto.

SIMONYAN, K.; ZISSERMAN, A. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, v. 1, 06 2014. Nenhuma citação no texto.

Singh, B. et al. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 1961–1970. Nenhuma citação no texto.

SMITH, C. P. A phoneme detector. *The Journal of the Acoustical Society of America*, v. 23, n. 4, p. 446–451, 1951. Disponível em: <https://doi.org/10.1121/1.1906786>. Nenhuma citação no texto.

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, NY, USA: ACM, 2005. (MULTIMEDIA '05), p. 399–402. ISBN 1-59593-044-2. Disponível em: <http://doi.acm.org/10.1145/1101149.1101236>. Nenhuma citação no texto.

Song, G.; Wang, S.; Tian, Q. Fusing feature and similarity for multimodal search. In: *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*. [S.l.: s.n.], 2015. p. 787–791. Nenhuma citação no texto.

SOOMRO, K.; ZAMIR, A. R.; SHAH, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. Disponível em: <http://arxiv.org/abs/1212.0402>. Nenhuma citação no texto.

SPERDUTI, A.; STARITA, A. Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, v. 8, n. 3, p. 714–735, 1997. Nenhuma citação no texto.

STEVENS, K. N. Autocorrelation analysis of speech sounds. *The Journal of the Acoustical Society of America*, v. 22, n. 6, p. 769–771, 1950. Disponível em: <https://doi.org/10.1121/1.1906687>. Nenhuma citação no texto.

SU, C.-W. et al. Motion flow-based video retrieval. *Multimedia, IEEE Transactions on*, v. 9, p. 1193 – 1201, 11 2007. Nenhuma citação no texto.

SULTANA, M.; GAVRILOVA, M. A content based feature combination method for face recognition. In: _____. *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013*. Heidelberg: Springer International Publishing, 2013. p. 197–206. ISBN 978-3-319-00969-8. Disponível em: <https://doi.org/10.1007/978-3-319-00969-8_19>. Nenhuma citação no texto.

SURESHA, M.; KUPPA, S.; RAGHUKUMAR, D. S. A study on deep learning spatiotemporal models and feature extraction techniques for video understanding. *International Journal of Multimedia Information Retrieval*, v. 9, n. 2, p. 81–101, Jun 2020. ISSN 2192-662X. Disponível em: <https://doi.org/10.1007/s13735-019-00190-x>. Nenhuma citação no texto.

SWAIN, M. J.; BALLARD, B. H. Color indexing. *International Journal of Computer Vision*, v. 7, n. 1, p. 11–32, 1991. Nenhuma citação no texto.

TAN, M.; LE, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv, 2019. Disponível em: <https://arxiv.org/abs/1905.11946>. Nenhuma citação no texto.

TANISIK, G.; ZALLUHOGLU, C.; IKIZLER-CINBIS, N. Facial descriptors for human interaction recognition in still images. *Pattern Recognition Letters*, v. 73, p. 44–51, 2016. ISSN 0167-8655. Nenhuma citação no texto.

TAO, B.; DICKINSON, B. W. Texture recognition and image retrieval using gradient indexing. *JVCIR*, v. 11, n. 3, p. 327–342, 2000. ISSN 1047-3203. Nenhuma citação no texto.

TORRES, R. da S.; FALCÃO, A. X. Content-Based Image Retrieval: Theory and Applications. *Revista de Informática Teórica e Aplicada*, v. 13, n. 2, p. 161–185, 2006. Nenhuma citação no texto.

TORRES, R. da S.; FALCÃO, A. X. Contour Salience Descriptors for Effective Image Retrieval and Analysis. *Image and Vision Computing*, v. 25, n. 1, p. 3–13, 2007. Nenhuma citação no texto.

TRAN, D. et al. A closer look at spatiotemporal convolutions for action recognition. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 6450–6459. Nenhuma citação no texto.

Tsau, E.; Kim, S.; Kuo, C. . J. Environmental sound recognition with celp-based features. In: *ISSCS 2011 - International Symposium on Signals, Circuits and Systems*. [S.l.: s.n.], 2011. p. 1–4. Nenhuma citação no texto.

Tzanetakis, G.; Cook, P. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, v. 10, n. 5, p. 293–302, 2002. Nenhuma citação no texto.

V K, J.; GURU, D. S.; Y H, S. K. Deep learning for retrieval of natural flower videos. *Procedia Computer Science*, v. 132, p. 1533 – 1542, 2018. ISSN 1877-0509. International Conference on Computational Intelligence and Data Science. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050918308494>. Nenhuma citação no texto.

VAIDYA, S.; SHAH, K. Audio denoising, recognition and retrieval by using feature vectors. In: *Journal of Computer Engineering (IOSR-JCE)*. [S.l.: s.n.], 2014. v. 16, n. 2, p. 107–112. Nenhuma citação no texto.

VALEM, L. P.; PEDRONETTE, D. C. G. Unsupervised similarity learning through cartesian product of ranking references for image retrieval tasks. *SIBGRAPI*, 2016. Nenhuma citação no texto.

VALEM, L. P.; PEDRONETTE, D. G. Unsupervised distance learning framework for multimedia retrieval. In: *ACM International Conference on Multimedia Retrieval*. [S.l.: s.n.], 2017. Nenhuma citação no texto.

VALEM L. P., P. D. C. G. Unsupervised similarity learning through cartesian product of ranking references for image retrieval tasks. In: *SIBGRAPI Conference on Images and Patterns*. [S.l.: s.n.], 2016. Nenhuma citação no texto.

VARAMESH, A. et al. *Self-Supervised Ranking for Representation Learning*. 2020. Nenhuma citação no texto.

VELIčKOVIć, P. et al. Deep graph infomax. In: *International Conference on Learning Representations*. [S.l.: s.n.], 2019. Nenhuma citação no texto.

WANG, J. et al. Learning context-sensitive similarity by shortest path propagation. *Pattern Recognition*, v. 44, n. 10, p. 2367 – 2374, 2011. ISSN 0031-3203. Semi-Supervised Learning for Visual Content Analysis and Understanding. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320311000549>. Nenhuma citação no texto.

Wang, J. et al. Learning fine-grained image similarity with deep ranking. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 1386–1393. Nenhuma citação no texto.

Wang, M. et al. Multimodal graph-based reranking for web image search. *IEEE Transactions on Image Processing*, v. 21, n. 11, p. 4649–4661, 2012. Nenhuma citação no texto.

WANG, M.; SONG, T. Remote sensing image retrieval by scene semantic matching. *IEEE Transactions on Geoscience and Remote Sensing*, v. 51, n. 5, p. 2874–2886, May 2013. ISSN 0196-2892. Nenhuma citação no texto.

WIGGERS, K. L. et al. Deep learning approaches for image retrieval and pattern spotting in ancient documents. *CoRR*, abs/1907.09404, 2019. Disponível em: <http://arxiv.org/abs/1907.09404>. Nenhuma citação no texto.

WU, Z. et al. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, v. 32, n. 1, p. 4–24, 2021. Nenhuma citação no texto.

XU, J. et al. Msr-vtt: A large video description dataset for bridging video and language. In: . [S.l.]: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2016. Nenhuma citação no texto.

Xue-Yan Zhao; Fei Wu; Jie Lin. Audio retrieval: based on unsupervised learning approach. In: *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826)*. [S.l.: s.n.], 2004. v. 3, p. 1625–1628 vol.3. Nenhuma citação no texto.

YAJIMA, C.; NAKANISHI, Y.; TANAKA, K. Querying video data by spatio-temporal relationships of moving object traces. In: _____. *Visual and Multimedia Information Management: IFIP TC2/WG2.6 Sixth Working Conference on Visual Database Systems May 29–31, 2002 Brisbane, Australia*. Boston, MA: Springer US, 2002. p. 357–371. ISBN 978-0-387-35592-4. Disponível em: <https://doi.org/10.1007/978-0-387-35592-4_25>. Nenhuma citação no texto.

YAN, R.; HAUPTMANN, A. A review of text and image retrieval approaches for broadcast news video. *Inf. Retr.*, v. 10, p. 445–484, 09 2007. Nenhuma citação no texto.

Yang, H.; Meinel, C. Content based lecture video retrieval using speech and video text information. *IEEE Transactions on Learning Technologies*, v. 7, n. 2, p. 142–154, 2014. Nenhuma citação no texto.

Yang, X. et al. Comparative study on voice activity detection algorithm. In: *2010 International Conference on Electrical and Control Engineering*. [S.l.: s.n.], 2010. p. 599–602. Nenhuma citação no texto.

YAO, B. et al. Human action recognition by learning bases of action attributes and parts. In: *Int. Conf. Computer Vision (ICCV'11)*. [S.l.: s.n.], 2011. p. 1331–1338. Nenhuma citação no texto.

Yu-Fei Ma; Hong-Jiang Zhang. Motion texture: a new motion based video representation. In: *Object recognition supported by user interaction for service robots*. [S.l.: s.n.], 2002. v. 2, p. 548–551 vol.2. Nenhuma citação no texto.

YU, J. et al. Feature integration analysis of bag-of-features model for image retrieval. *Neurocomputing*, v. 120, n. Supplement C, p. 355 – 364, 2013. ISSN 0925-2312. Image Feature Detection and Description. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231213003020>. Nenhuma citação no texto.

YUE, J. et al. Content-based image retrieval using color and texture fused features. *Mathematical and Computer Modelling*, v. 54, n. 3, p. 1121 – 1127, 2011. ISSN 0895-7177. Mathematical and Computer Modeling in agriculture (CCTA 2010). Disponível em: <http://www.sciencedirect.com/science/article/pii/S0895717710005352>. Nenhuma citação no texto.

ZENG, Y. et al. Multi-modal relational graph for cross-modal video moment retrieval. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2021. p. 2215–2224. Nenhuma citação no texto.

ZHANG, S. et al. Query specific rank fusion for image retrieval. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. [S.l.: s.n.], 2012. v. 37, p. 660–673. ISBN 978-3-642-33708-6. Nenhuma citação no texto.

ZHANG, W.; QIN, Z.; WAN, T. Image scene categorization using multi-bag-of-features. In: *2011 International Conference on Machine Learning and Cybernetics*. [S.l.: s.n.], 2011. v. 4, p. 1804–1808. ISSN 2160-133X. Nenhuma citação no texto.

ZHANG, X.-Y. et al. Learning transferable self-attentive representations for action recognition in untrimmed videos with weak supervision. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 33, p. 9227–9234, 07 2019. Nenhuma citação no texto.

Zhang, Y. et al. Acoustic scene classification using deep audio feature and blstm network. In: *2018 International Conference on Audio, Language and Image Processing (ICALIP)*. [S.l.: s.n.], 2018. p. 371–374. Nenhuma citação no texto.

ZHOU, X.; DEPEURSINGE, A.; MULLER, H. Information fusion for combining visual and textual image retrieval. In: *2010 20th International Conference on Pattern Recognition*. [S.l.: s.n.], 2010. p. 1590–1593. Nenhuma citação no texto.

Zhu, Q.; Shyu, M. Sparse linear integration of content and context modalities for semantic concept retrieval. *IEEE Transactions on Emerging Topics in Computing*, v. 3, n. 2, p. 152–160, 2015. Nenhuma citação no texto.