



UNIVERSIDADE ESTADUAL PAULISTA
“JÚLIO DE MESQUITA FILHO”
Câmpus de Rio Claro

Instituto de Geociências e Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Ademir Moreno Junior

Seleção e agregação de dupla camada para listas ranqueadas em recuperação de imagens por conteúdo

Rio Claro - SP

2024

UNIVERSIDADE ESTADUAL PAULISTA

“Júlio de Mesquita Filho”

Instituto de Geociências e Ciências Exatas

Câmpus de Rio Claro

Ademir Moreno Junior

**Seleção e agregação de dupla camada para listas
ranqueadas em recuperação de imagens por conteúdo**

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.

Orientador: Prof. Dr. Daniel Carlos
Guimarães Pedronette

Rio Claro - SP

2024

M843s Moreno Junior, Ademir

Seleção e agregação de dupla camada para listas ranqueadas em recuperação de imagens por conteúdo / Ademir Moreno Junior. -- Rio Claro, 2024

97 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Instituto de Geociências e Ciências Exatas, Rio Claro

Orientador: Daniel Carlos Guimarães Pedronette

1. Recuperação de imagens baseado em conteúdo. 2. Sistemas de recuperação da informação. 3. Ranqueamento. 4. Agregação de listas ranqueadas. 5. Aprendizado de máquina. I. Título.

UNIVERSIDADE ESTADUAL PAULISTA
“Júlio de Mesquita Filho”
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

Ademir Moreno Junior

Seleção e agregação de dupla camada para listas ranqueadas em recuperação de imagens por conteúdo

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 (Orientador)
Instituto de Geociências e Ciências Exatas (IGCE)
Universidade Estadual Paulista - UNESP
- Prof^a. Dr^a. Veronica Oliveira de Carvalho
Instituto de Ciências Matemáticas e de Computação (IGCE)
Universidade de São Paulo - USP
- Prof. Dr. Guilherme Palermo Coelho
Universidade Estadual de Campinas - UNICAMP

Conceito: Aprovado.

Rio Claro (SP), 26 de setembro de 2024.

Agradecimentos

Agradeço primeiramente a Deus pela vida, toda saúde e disposição.

Aos meus amigos e familiares, pelo apoio incondicional.

Ao meu orientador, pela orientação, apoio e confiança.

À Fundação de Amparo à Pesquisa do Estado de São Paulo pelo apoio financeiro.

Agradeço à universidade e a todo o corpo docente do programa de pós-graduação.

Aos amigos e colegas, pelo incentivo e apoio.

processos nº #2018/15597-6 e #2022/07349-8, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)

“Em primeiro lugar vem a dedicaç o, depois a habilidade.”
(Leonardo da Vinci)

Resumo

Em decorrência de amplos avanços tecnológicos nos processos de aquisição e armazenamento de imagens, houve um massivo crescimento de coleções multimídia nas últimas décadas. Tal crescimento desencadeou uma demanda premente por ferramentas automáticas para mitigar a sobrecarga de informações. O uso de técnicas de recuperação de imagens capazes de levar em consideração o conteúdo visual tornou-se essencial neste contexto. Os sistemas de Recuperação de Imagens Baseados em Conteúdo (CBIR) consolidaram-se como uma solução relevante e representam um campo de pesquisa extremamente ativo com diversificadas aplicações. Tais sistemas comumente baseiam-se na representação de imagens por meio de um processo de extração de características e na comparação das representações para a obtenção de listas ranqueadas. Contudo, há numerosas abordagens diferentes para extração de características. Apesar da revolução da área suportada por métodos baseados em aprendizado profundo, a disponibilidade de diferentes características geradas por diferentes extratores se mantém. Assim, a obtenção de resultados mais eficazes obtidos por meio da combinação de listas ranqueadas continua uma solução atrativa, uma vez que diferentes abordagens de extração de características e ranqueamento originam listas ranqueadas distintas. Contudo, a combinação de listas ranqueadas em cenários não supervisionados apresenta vários desafios. Entre eles, pode-se destacar: a dificuldade em selecionar os extratores de características e os métodos utilizados para combinação em cenários de ausência de rótulos. Com o propósito de contribuir com esse campo de pesquisa, esse trabalho apresenta uma proposta de metodologia de seleção e agregação multinível de ranqueadores baseados em diferentes extratores de características. A proposta deste trabalho aborda um cenário de agregação de ranqueadores em um contexto não supervisionado, estruturado em duas camadas. A primeira camada baseia-se em medidas de estimativa de eficácia para selecionar ranqueadores individuais, os quais são combinados em pares de forma sistemática. Na segunda camada, os pares de ranqueadores gerados na etapa anterior são selecionados e combinados, resultando em uma lista ranqueada final. Os resultados obtidos em avaliação experimental conduzida em três *datasets* de diferentes tamanhos demonstram que o método proposto apresenta resultados que superam metodologias já estabelecidas na área.

Palavras-chave: Recuperação de Informações, Recuperação de Imagens por Conteúdo, Seleção e Combinação de Características, Seleção e Combinação de Listas, Reconhecimento de Padrões, Métodos de Aprendizado Não Supervisionado, rerranqueamento, multinível, dupla camada.

Abstract

Due to significant technological advances in the processes of image acquisition and storage, there has been a massive growth in multimedia collections in recent decades. This growth has triggered an urgent demand for automatic tools to mitigate information overload. The use of image retrieval techniques capable of considering visual content has become essential in this context. Content-Based Image Retrieval (CBIR) systems have emerged as a relevant solution and represent an extremely active research field with diverse applications. These systems are commonly based on the representation of images through a feature extraction process and the comparison of these representations to obtain ranked lists. However, there are numerous different approaches to feature extraction. Despite the revolution in the field supported by deep learning-based methods, the availability of different features generated by various extractors remains. Thus, achieving more effective results through the combination of ranked lists continues to be an attractive solution, as different feature extraction and ranking approaches produce distinct ranked lists. Nevertheless, the combination of ranked lists in unsupervised scenarios presents several challenges. Among them, the difficulty in selecting feature extractors and methods for combination in the absence of labels stands out. In order to contribute to this research field, this work presents a proposal for a multilevel selection and aggregation methodology for rankers based on different feature extractors. The proposal addresses a scenario of ranker aggregation in an unsupervised context, structured in two layers. The first layer is based on efficacy estimation metrics to select individual rankers, which are systematically combined in pairs. In the second layer, the ranker pairs generated in the previous step are selected and combined, resulting in a final ranked list. The results obtained from experimental evaluation conducted on three datasets of different sizes demonstrate that the proposed method outperforms well-established methodologies in the field.

Keywords: Information Retrieval, Content-Based Image Retrieval, Feature Combination, Feature Selection, Ranking Combination, Ranking Selection, Pattern Recognition, Statistics, Unsupervised Learning, Re-ranking, multi-level, layers.

Lista de ilustrações

Figura 1 – Contextualização do processo de extração de características.	20
Figura 2 – Arquitetura típica de um sistema CBIR.	21
Figura 3 – Ilustração da questão da Lacuna Semântica.	22
Figura 4 – Ilustração de uma imagem em tons de cinza e seu respectivo histograma.	23
Figura 5 – Ilustração de uma imagem colorida e seus histogramas de cor para cada canal.	24
Figura 6 – Ilustração de uma matriz LBP.	25
Figura 7 – Ilustração de uma matriz HoG.	26
Figura 8 – Ilustração da esquemática de um sistema baseado em CBIR.	35
Figura 9 – Ilustração do fluxo de combinação de um sistema baseado em <i>Early Fusion</i>	35
Figura 10 – Ilustração do fluxo de combinação de um sistema baseado em <i>Late Fusion</i>	36
Figura 11 – Agregação com cadeia de Markov, executando a agregação das listas.	39
Figura 12 – Ilustração do grafo de vizinhança com ocorrência de vizinhos recíprocos.	41
Figura 13 – Esquemática de um sistema de agregação baseado em grafos.	46
Figura 14 – Diagrama de agregação multinível baseada em hierarquia.	48
Figura 15 – Arquitetura do modelo de fusão multinível.	49
Figura 16 – Ilustração Método USRF.	52
Figura 17 – Arquitetura do modelo VGG-19.	53
Figura 18 – Fluxo da abordagem baseada em pesos.	55
Figura 19 – Fluxo de agregação de metavértices.	56
Figura 20 – Passos iniciais do UDLF.	61
Figura 21 – Organização geral da metodologia de agregação de ranqueadores multinível.	63
Figura 22 – Ilustração do processo de seleção de ranqueadores.	65
Figura 23 – Agregação de ranqueadores par-a-par.	66
Figura 24 – Seleção de ranqueadores combinados.	68
Figura 25 – Ilustração do processo de agregação final.	69
Figura 26 – Amostra de elementos do conjunto Oxford <i>Flowers</i> -17.	72
Figura 27 – Amostra de elementos do conjunto Corel5K.	72
Figura 28 – Amostra de elementos do conjunto Cub200.	73
Figura 29 – Gráfico de relação MAP e <i>Authority</i> para o <i>dataset flowers</i> -17.	76
Figura 30 – Gráfico de relação MAP e <i>Authority</i> para o <i>dataset Corel5K</i>	76
Figura 31 – Gráfico de relação MAP e <i>Authority</i> para o <i>dataset Cub200</i>	77
Figura 32 – Gráfico de comparação de combinações e linha do melhor descritor isolado para o <i>Flowers</i> -17.	81

Figura 33 – Gráfico de comparação de combinações e linha do melhor descritor isolado para o Corel5K.	81
Figura 34 – Gráfico de comparação de combinações e linha do melhor descritor isolado para o Cub200.	82
Figura 35 – Resultados visuais para as imagens 2178 e 1852 do <i>dataset</i> Corel5K. . .	84
Figura 36 – Resultados visuais para as imagens 1178, 1200 e 2157 do <i>dataset</i> Corel5K.	85

Lista de tabelas

Tabela 1	– Exemplos de métricas de distância.	20
Tabela 2	– Métodos lineares de combinação (COMB).	37
Tabela 3	– Resultados isolados de cada descritor para o <i>dataset</i> Flowers-17 ordenado pelo valor da coluna Borda Rank.	75
Tabela 4	– Resultados isolados de cada descritor para o <i>dataset</i> Corel5K ordenado pelo valor da coluna Borda Rank.	75
Tabela 5	– Resultados isolados de cada descritor para o <i>dataset</i> Cub200 ordenado pelo valor da coluna Borda Rank.	75
Tabela 6	– Resultados <i>dataset</i> Flowers-17 ordenados pelo Borda Rank	78
Tabela 7	– Resultados <i>dataset</i> Corel5k ordenados pelo Borda Rank	79
Tabela 8	– Resultados <i>dataset</i> Cub200 ordenados pelo Borda Rank	80
Tabela 9	– Resultados obtidos com a aplicação do método UDLA para os <i>datasets</i> e a comparação com o método USRF.	82

Lista de abreviaturas e siglas

ACC	<i>Color Autocorrelator</i>
BIC	<i>Border/Interior Classification</i>
BoW	<i>Bag of Words</i>
BoVW	<i>Bag of Visual Words</i>
BN	<i>Batch Normalization</i>
BRIEF	<i>Binary Robust Independent Elementary Feature</i>
CBIR	<i>Content-Based Image Retrieval</i>
CFD	<i>Contour Features Descriptor</i>
CNN	<i>Convolutional Neural Network</i>
CODRA	<i>Cosine Distance for Rank Aggregation</i>
CS	<i>Contour Saliency</i>
DBN	<i>Deep Belief Network</i>
DDBTC	<i>Dot-Diffused Block Truncation Coding</i>
DoG	<i>Difference of Gaussian</i>
DNN	<i>Deep Neural Network</i>
FAST	<i>Features from Accelerated Segment Test</i>
FGIR	<i>Fine-grained image retrieval</i>
FGSBIR	<i>fine-grained sketch-based image retrieval</i>
GCH	<i>Global Color Histogram</i>
GLCM	<i>Grey Level Co-occurrence Matrices</i>
GN	<i>Group Normalization</i>
HoG	<i>Histogram of Oriented Gradient</i>
HSI	<i>Hue, Saturation and intensity</i>

IoU	<i>Intersection over Union</i>
LB	<i>Lovasz Bregman</i>
LBP	<i>Local Binary Pattern</i>
MFFN	<i>Multiple-level Feature Fusion Network</i>
MRA	<i>Median Rank Aggregation</i>
NDCG	<i>Normalized Discounted Cumulative Gain</i>
ORB	<i>Oriented FAST Rotate BRIEF</i>
RBM	<i>Restricted Boltzmann Machine</i>
SCDA	<i>Selective Convolutional Descriptor Aggregation</i>
SIFT	<i>Scale-Invariant Feature Transform</i>
SURF	<i>Speeded Up Robust Features</i>
SS	<i>Segment Saliences</i>
SVM	<i>Support Vector Machine</i>
TNN	<i>Transformer Neural Network</i>
UDLF	<i>Unsupervised Distance Learning Framework</i>
ULARA	<i>Unsupervised Learning Algorithm for Rank Aggregation</i>
VLAD	<i>Vector of Locally Aggregated Descriptors</i>
VPR	<i>Visual Place Recognition</i>
FV	<i>Features vector</i>
TBIR	<i>Text-Based Image Retrieval</i>

Sumário

1	INTRODUÇÃO	15
2	RECUPERAÇÃO DE IMAGENS BASEADA EM CONTEÚDO . . .	19
2.1	Técnicas de Extração de Características	22
2.2	Métricas de Avaliação de Resultados	32
3	ABORDAGENS DE COMBINAÇÃO NÃO SUPERVISIONADAS . .	34
3.1	Definição Formal	34
3.2	Técnicas de Agregação	34
3.3	Técnicas de Agregação de Listas Ranqueadas	36
3.4	Técnicas de Seleção de Descritores Não Supervisionada	40
3.4.1	Estimativas de Eficácia	40
3.4.2	Métricas de Correlação	42
4	ABORDAGENS RECENTES	45
4.1	Abordagens baseadas em Grafo	45
4.2	Abordagens de Agregação Hierárquica de Características e Abordagens de Agregação multinível	47
4.3	Abordagens baseadas em Redes Neurais	52
4.4	Abordagens de Outras Áreas de Agregação com Potencial para Aplicação em Recuperação de Imagens	54
4.5	Outras Abordagens	57
4.6	Bibliotecas de Agregação Não Supervisionada de Listas Ranqueadas	59
5	METODOLOGIA DE AGREGAÇÃO NÃO SUPERVISIONADA EM DUPLA CAMADA (UDLA)	62
5.1	Seleção de Ranqueadores	63
5.2	Agregação das Listas Ranqueadas dos Descritores	66
5.3	Seleção de Melhores Combinações	67
5.4	Agregação de Combinações	69
6	AVALIAÇÃO EXPERIMENTAL	71
6.1	<i>Datasets</i>	71
6.1.1	Flowers-17	71
6.1.2	Corel5K	72
6.1.3	Cub200	73
6.2	Protocolo Experimental, Parâmetros e Ranqueadores	73

6.2.1	Resultados Isolados	74
6.3	Resultados de Combinação	77
6.4	Resultados Visuais	83
7	CONCLUSÃO	87
	REFERÊNCIAS	89

1 Introdução

Ao longo da história, a humanidade tem se empenhado em preservar e recuperar informações e conhecimentos. No início, esses conhecimentos eram registrados por meio de desenhos e documentos escritos. Estruturas baseadas em bibliotecas, como a famosa Biblioteca de Alexandria, foram construídas visando garantir um armazenamento adequado para livros, papiros e documentos. Com o aumento contínuo no volume desses documentos, houve uma demanda urgente por um sistema que permitisse a localização rápida e eficiente desses itens. Portanto, os índices foram desenvolvidos, permitindo a identificação e localização de cada documento. É importante ressaltar que esse método era realizado manualmente nos primórdios de sua concepção (BAEZA-YATES; RIBEIRO-NETO, 2013).

A ampliação das coleções no decorrer do tempo tornou o processo de indexação muito complexo para execução manual. Com a popularização dos computadores o processo de indexação tornou-se algo passível de automatização, possibilitando a localização de itens em coleções extensas. Com a chegada dos computadores, além da indexação automática, agregou-se um crescimento gigantesco na quantidade de informação armazenada. Assim, os conceitos de Recuperação da Informação (RI) aplicados para recuperação de livros e documentos físicos foram adaptados e aprimorados para recuperação de informações digitais, que eram geradas aceleradamente associados ao crescimento de pessoas com acesso à *internet* (KOBAYASHI; TAKEDA, 2000).

Além de documentos textuais, as pinturas e conteúdos visuais representam aspectos relevantes da vida humana. Desde os primórdios, os humanos registraram informações em pinturas nas paredes das cavernas, conhecidas como pinturas rupestres, sobre como se proteger de predadores e armazenar alimentos. O uso de imagens pelo ser humano desempenha um papel fundamental em diversificados domínios. Desde a invenção do raio-x, no final dos anos 1895, o uso de imagens tornou-se intrínseco no cotidiano do campo da saúde (PINTO, 2008; AFIFI; ASHOUR, 2012).

Com a evolução dos computadores e a expansão da mineração de dados em quantidades massivas, a Recuperação da Informação (RI) tornou-se uma área em destaque, assim como a aplicação de metodologias baseadas em aprendizado de máquina para uma recuperação eficaz e eficiente do vasto repositório de conteúdos disponíveis (SALMAN; AL-ZAATREH; SULIEMAN, 2022). Tendo a evolução computacional progredido amplamente em reduzidos períodos, mais indivíduos tiveram oportunidade de acesso à tecnologia. Em decorrência, logo uma quantidade maior de conteúdos multimídia começou a ser criada e armazenada, e assim como os textos, tais arquivos necessitavam de uma metodologia de recuperação capaz de buscar os objetos digitais multimídia considerando seu conteúdo.

Nesse cenário, as metodologias de RI foram aprimoradas para contemplar a recuperação desse tipo de mídia, dando origem aos sistemas de Recuperação de Imagem baseado em Conteúdo (CBIR) (VALEM, 2019).

Entre os anos de 1994 e 2000, as metodologias de CBIR foram introduzidas como um primeiro passo nas pesquisas para metodologias de recuperação de imagem por conteúdo. A motivação principal para essas metodologias foi a percepção dos pesquisadores de que era possível abstrair características de imagens, possibilitando obter resultados de pesquisa eficazes conforme as necessidades dos usuários ao considerar tais características (DATTA et al., 2008). Para a recuperação e análise de imagens, um dos processos fundamentais consiste em extrair uma determinada característica visual, como cor, forma, textura, entre outras que podem incluir combinações dessas. Desde a última década, o processo de extração de características passou por importantes revoluções. Ganharam espaço na academia e na indústria abordagens baseadas em redes neurais convolucionais treinadas por meio de transferência de aprendizado. Mais recentemente, modelos baseados em *Transformers* têm atingido resultados promissores e atraído muita atenção na literatura.

Neste cenário, diversas abordagens foram propostas para extração de características. A ResNet é um exemplo de rede convolucional que têm recebido considerável atenção, tendo como ponto-chave o conceito de conexões residuais, que permitem que os gradientes sejam passados diretamente por camadas, pulando uma ou mais delas. Essas conexões ajudam a evitar o problema do desaparecimento ou explosão do gradiente durante o treinamento, o que ocorre em redes profundas tradicionais. Isso permite que a ResNet tenha centenas ou até milhares de camadas, mantendo a capacidade de aprendizado e melhorando o desempenho em tarefas como classificação de imagens (HE et al., 2016; HE; LIU; TAO, 2020). Um exemplo de rede *Transformers* é a *Visual Transformers* (ViT). Essas redes possuem campos receptivos amplos e facilitam a interação entre vários recursos semânticos de alto nível, resultando em melhorias de desempenho em comparação com redes convolucionais tradicionais (LIU et al., 2023a). Outra rede *transformer* que vem se destacando é a *Distillation with No Labels Version 2* (DINOv2), que apresenta um diferencial dos modelos anteriores principalmente em sua abordagem de aprendizado autossupervisionado, arquitetura, métodos de treinamento e capacidade de produzir recursos visuais robustos sem supervisão (OQUAB et al., 2023).

Contudo, seja considerando abordagens tradicionais ou baseadas em aprendizado profundo, há um amplo espectro de abordagens diferentes que podem ser utilizados para extração de características de imagens. Frequentemente, a aplicação de apenas um extrator pode ser insuficiente para representar diferentes aspectos da imagem, ocasionando resultados de lista ranqueadas de baixa eficácia. Neste cenário, uma solução promissora consiste em explorar de forma conjunto os resultados de diferentes abordagens de extração de características. Tais técnicas são denominadas de técnicas de fusão ou combinação e

são comumente divididas em categorias denominadas de *early fusion* e *late fusion* (YAO et al., 2021). Alguns autores acrescentam uma terceira classificação, denominada *intermediate fusion*, a qual está geralmente relacionada ao uso de classificadores em etapa intermediária (IONESCU et al., 2014).

Neste trabalho, um enfoque maior será conferido às técnicas de *late fusion*, modelo de fusão em que se obtém uma lista ranqueada individual para cada extrator que servirá de base para a aplicação das técnicas de agregação de listas (*rank aggregation*). Agregação de listas ranqueadas é um processo tradicional de combinação de preferências ordenadas geradas por mais de um ranqueador (SCULLEY, 2007) e é aplicada visando aprimorar os resultados obtidos mediante listas ranqueadas obtidas por meio de diferentes descritores (LIU, ; SCHALEKAMP; ZUYLEN, 2009; LIU et al., 2007). Fundamentais para a metodologia proposta, tais métodos permitem combinar as listas ranqueadas obtidas a partir de duas ou mais características extraídas. Devido à variedade, o posicionamento do elemento dependerá de sua repetição e ranqueamento nas listas resultantes, produzindo assim uma nova lista com um *ranking* que tende a ser mais precisa ao ser comparada as listas de características individuais (BAŁCHANOWSKI; BORYCZKA, 2022).

Embora frequentemente a agregação leve a ganhos de eficácia, há vários desafios associados, especialmente em cenários de ausência de rótulos. Em aplicações não supervisionadas, em que não há informações sobre a eficácia das características, combinar ranqueadores de menor eficácia pode levar a resultados menos eficazes. Assim, um desafio fundamental na tarefa consiste na seleção não supervisionada dos ranqueadores. Apesar dos avanços anteriores realizados nessa direção (VALEM; PEDRONETTE, 2020b; VALEM; PEDRONETTE, 2020a), há dificuldades na seleção das características em uma etapa única, devido ao grande número de combinações possíveis.

O objetivo principal deste estudo consiste na proposta de uma abordagem multinível para seleção e combinação de características visuais. Como resultado do estudo, é proposta uma metodologia de seleção e agregação não supervisionada em duas camadas. Na primeira camada, são utilizadas medidas de estimativa de eficácia baseadas em ranqueamento para estimar os m ranqueadores mais eficazes. Para os ranqueadores selecionados, aplica-se uma abordagem combinatória para definir os pares de ranqueadores combinados. As melhores combinações de pares obtidas na primeira camada são agregadas na segunda camada, obtendo-se uma nova lista ranqueada final. A metodologia proposta procura aumentar a eficácia ao selecionar e combinar os pares em oposição a fazer uma combinação direta dos ranqueadores. Adicionalmente, a abordagem proposta utiliza métodos contextuais de agregação de listas distintos nas duas camadas, com foco em precisão na primeira camada e revocação na segunda camada.

Para uma melhor compreensão são discutidos fundamentos relacionados aos modelos de recuperação de imagens baseados em conteúdo (CBIR), descritores locais e globais,

bem como métodos de agregação tradicionais e recentes. Além disso, medidas para estimar a eficácia e técnicas de seleção não supervisionada também são abordados. O restante do texto é organizado como descrito a seguir:

- No Capítulo 2 são abordadas as técnicas de recuperação de imagem baseadas em conteúdo (CBIR). No conceito de recuperação de imagem baseada em conteúdo serão abordadas as técnicas de extração de características na Seção 2.1, incluindo uma revisão sobre descritores baseados em cor na Seção 2.1, baseados em textura, baseados em gradiente, locais, baseados em redes neurais convolucionais e em redes neurais *Transformers*.

Na Seção 2.2 serão abordadas as métricas de avaliação de resultados para algoritmos de recuperação da informação;

- O Capítulo 3 descreve as abordagens de combinação não supervisionadas, abordando as técnicas de agregação de listas ranqueadas na Seção 3.3 e as técnicas de seleção não supervisionada na Seção 3.4;
- No Capítulo 4 serão apresentadas as abordagens recentes referentes ao tema;
- No Capítulo 5 será apresentada a metodologia proposta para este trabalho, sendo abordadas suas etapas nas Seções de seleção dos descritores 5.1, combinação dos descritores 5.2, seleção de listas ranqueadas combinadas 5.3 e por fim o processo de agregação final na Seção 5.4;
- No Capítulo 6 são apresentadas as coleções de dados utilizadas para os experimentos na Seção 6.1. Os resultados obtidos com a aplicação da metodologia proposta aos *datasets* selecionados são apresentados na Seção 6.3. Uma representação visual para os resultados é apresentada na Seção 6.4;
- Por fim, o Capítulo 7 apresenta a conclusão e discute direções para trabalhos futuros na área.

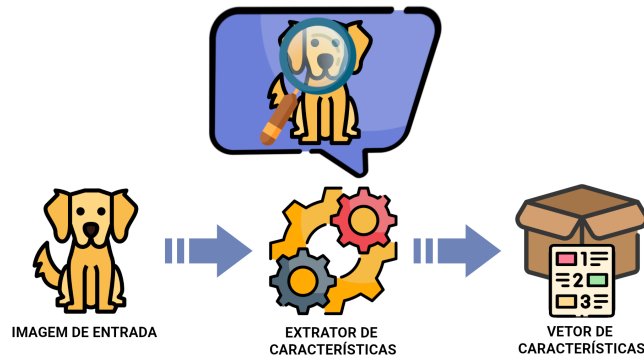
2 Recuperação de imagens Baseada em Conteúdo

A recuperação de imagens pode ser dividida em duas metodologias principais: a Recuperação de Imagem Baseada em Texto e a Recuperação de Imagem Baseada em Conteúdo. Essa divisão é essencial para compreender os aspectos fundamentais das metodologias de recuperação de imagem (ALKHAWLANI; ELMOGY; BAKRY, 2015).

Este capítulo terá como foco principal as estratégias de recuperação de imagens baseadas no conteúdo. Também serão discutidas as técnicas de extração de características, responsáveis pela obtenção das características para o cálculo de similaridade e criação das listas ranqueadas com base nessas características. Em seguida, essas listas serão utilizadas pelas técnicas de agregação de listas ranqueadas propostas.

As metodologias baseadas em conteúdo utilizam de consultas denominadas *Query By Image Content* (QBIC), tratando-se de uma forma automatizada de reconhecimento da imagem de consulta baseada características de imagens capazes de descrevê-las e retornando uma lista de imagens similares à imagem de consulta (TYAGI; RETRIEVAL, 2018).

O cérebro humano pode memorizar naturalmente características de um objeto, assim como possui a perícia para comparar se um segundo objeto possui uma quantidade relevante de características que permita a similaridade de ambos (CASTRO, 2006). Tal conceito pode ser assimilado como uma visão computacional bioinspirada, similar a proposta base dos métodos de recuperação de imagens baseada em conteúdo (CBIR), a qual é identificar as características do objeto de consulta (Imagem de entrada), extraí-las e em seguida comparar a similaridade entre as características extraídas e as características dos objetos contidos na *database* (NIXON; AGUADO, 2019), como visto na Figura 1.

Figura 1 – Contextualização do processo de extração de características.

Fonte: autor.

De maneira mais ampla, a recuperação de uma imagem usando um sistema baseado em CBIR envolve o envio de uma imagem de entrada (QBIC) para o processo de consulta. Durante esse processo, um vetor de características é derivado por meio de descritores, seguido pelo cálculo da similaridade através da avaliação da distância entre o vetor de consulta e os vetores de características armazenados na *database*. Geralmente, a distância Euclidiana é empregada para esse cálculo, embora outras métricas, conforme apresentadas na Tabela 1, possam ser utilizadas (DUBEY, 2021; HAMEED; ABDULHUSSAIN; MAHMMOD, 2021; SMEULDERS et al., 2000; ALKHAWLANI; ELMOGY; BAKRY, 2015; TYAGI; RETRIEVAL, 2018).

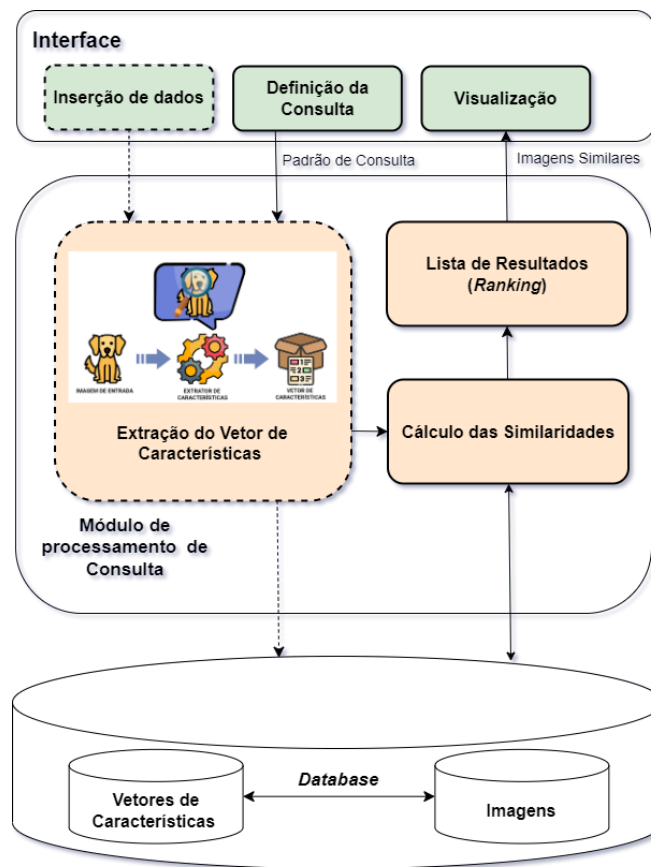
Tabela 1 – Exemplos de métricas de distância.

Distância	Equação
Chebyshev	$D(x, y) = \max_{i=1}^n x_i - y_i $
Euclidiana	$D(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$
Manhattan	$D(x, y) = \sum_{i=1}^n x_i - y_i $
Minkowsky	$D(x, y) = [\sum_{i=1}^n x_i - y_i ^p]^{1/p}$

Conforme descrito por Torres e Falcão, um sistema baseado em Recuperação de Imagens Baseada em Conteúdo (CBIR) possui duas divisões principais de funções: a entrada de dados e o processamento de consultas. O processo de entrada de dados é responsável pelo processamento das imagens para a extração de suas características, as quais são armazenadas juntamente com seus respectivos vetores de características. Geralmente, esse processo ocorre de forma *off-line* e os dados são armazenados em uma *database*. Também definem um descritor sendo denotado por uma tupla contendo a função de extração de características e a função de distância (TORRES; FALCÃO, 2006).

O procedimento de busca envolve a definição de uma imagem de consulta, que é então submetida ao processo de extração de características. Uma vez obtido esse vetor, é realizado o cálculo de similaridade entre a imagem de consulta e as imagens armazenadas, gerando um ranking de candidatos com base na semelhança de seus vetores de características. Esse processo de avaliação de similaridade ordena as imagens de acordo com sua proximidade com a imagem de consulta (TORRES; FALCÃO, 2006; HAMEED; ABDULHUSSAIN; MAHMMOD, 2021; DUBEY, 2021).

Figura 2 – Arquitetura típica de um sistema CBIR.



Fonte: Adaptado de (TORRES; FALCÃO, 2006).

Na Figura 2 é representado um sistema baseado em CBIR, onde se observa uma divisão em três partes, sendo a *interface* responsável pela interação com o usuário, o *Módulo de processamento de consultas* responsável pela extração de características, o cálculo de similaridade e a lista de resultados (*Ranking*), e a *database* onde são armazenadas as imagens e seus respectivos vetores de características.

2.1 Técnicas de Extração de Características

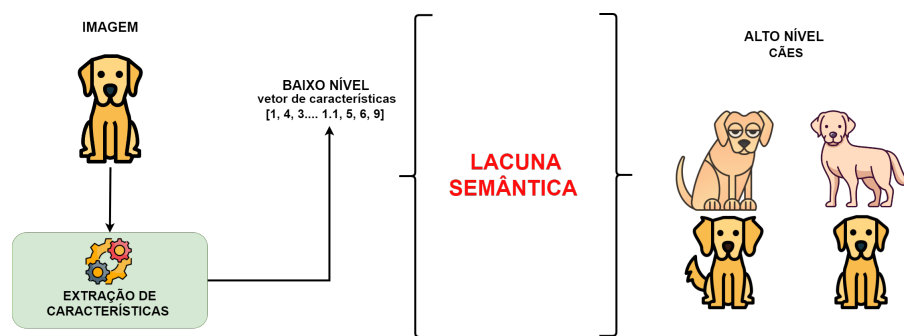
As técnicas de extração de características, também conhecidas como descritores, são algoritmos que codificam as características de uma imagem em um vetor numérico ordenado de características.

Se apenas os *pixels* de uma imagem, sem nenhuma técnica de pré-processamento, forem considerados, os resultados gerados serão pouco eficazes. Deste modo, os descritores visuais foram propostos como o objetivo de extrair características consideradas relevantes de uma imagem, e armazená-las em um vetor numérico denominado Vetor de características (DATTA et al., 2008; VALEM, 2019).

Em sistemas baseados em CBIR são geralmente utilizadas características de baixo nível, como cor, textura, forma e localização espacial para a recuperação. Vale ressaltar que existe uma divisão definida entre descritores globais, que aplicam um algoritmo a uma imagem na totalidade, e os descritores locais pensados para solucionar o problema de variância e aumento da eficácia das buscas por imagens por conteúdo (NIXON; AGUADO, 2019; TYAGI; RETRIEVAL, 2018; HAMEED; ABDULHUSSAIN; MAHMMOD, 2021).

A descrição de uma imagem representa um desafio amplamente discutido na literatura, especialmente ao considerar a lacuna semântica (*Semantic Gap*), que aborda a diferença entre o conceito de alto nível (a representação da imagem) e o conceito de baixo nível (o vetor de características). Isso é ilustrado na Figura 3, onde a limitação surge devido à representação da imagem apenas por um vetor numérico no vetor de características, incapaz de capturar totalmente o conteúdo, como a identificação específica de um cão (DATTA et al., 2008; TYAGI; RETRIEVAL, 2018; TORRES; FALCÃO, 2006; SMEULDERS et al., 2000).

Figura 3 – Ilustração da questão da Lacuna Semântica.



Fonte: autor.

Para tratar a questão da lacuna semântica, algumas soluções são propostas na literatura, geralmente atreladas ao uso de aprendizado profundo de máquina para ex-

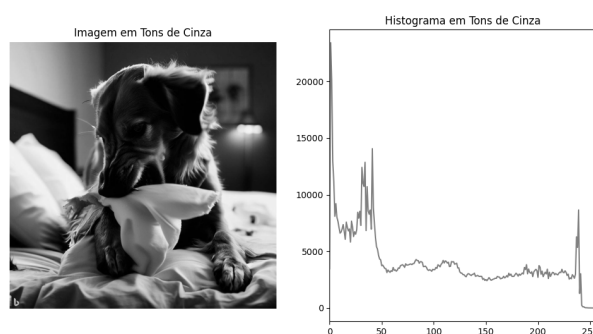
tração de características de alto nível com resultados que apresentam redução dessa lacuna (JABNOUN et al., 2022).

- **Descritores Baseados em Cor:**

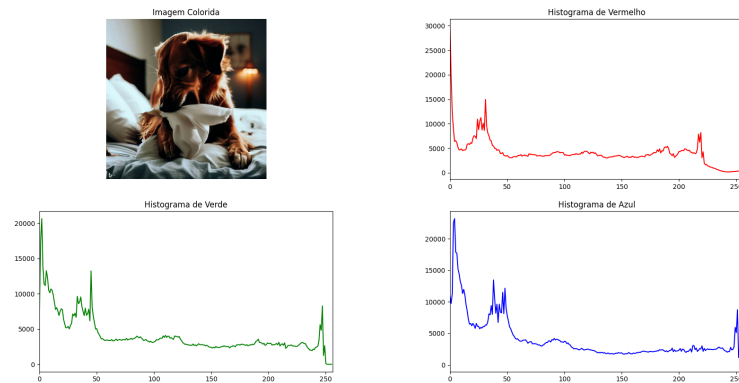
Descritores por cor são comumente utilizados em sistemas baseados em CBIR, uma vez que as escalas de cor são representadas por uma matriz tridimensional com variação de intensidade de 0 a 255 para cada dimensão, essa variação aplicada as dimensões representa o tom da cor de cada *pixel* da imagem. O padrão mais comum de cor é o RGB (*Red — Green — Blue*), mas existem outros como CMY (*Cyan — Magenta — Yellow*) e Lab, por exemplo. Vale ressaltar que imagens em tons de cinza possuem apenas uma dimensão com variação de 0 a 255 que implicará no nível de iluminação de cada *pixel* sendo 0 sem iluminação, ou seja, preto e 255 com iluminação máxima, ou seja, branco (TORRES; FALCÃO, 2006; NIXON; AGUADO, 2019).

Os descritores de cor computam um histograma para as cores de entrada, gerando um gráfico de frequência para cada tom, ou seja, quantos *pixels* da imagem possuem aquele valor. Na Figura 4 é ilustrado um histograma de uma imagem em tons de cinza, assim apenas a iluminação é analisada, em casos de imagens coloridas é gerado um histograma para cada dimensão das cores, como pode ser visto na Figura 5, mais comumente utilizado um canal vermelho, um verde e um azul, e então é contada a intensidade do tom de 0 a 255 em cada canal de cor (SMEULDERS et al., 2000; HAMEED; ABDULHUSSAIN; MAHMMOD, 2021; AFIFI; ASHOUR, 2012).

Figura 4 — Ilustração de uma imagem em tons de cinza e seu respectivo histograma.



Fonte: Adaptado de (MICROSOFT, 2023).

Figura 5 – Ilustração de uma imagem colorida e seus histogramas de cor para cada canal.

Fonte: Adaptado de (MICROSOFT, 2023).

Como descritores de cor pode-se utilizar de exemplo o *Global Color Histogram* (GCH) que propõe a normalização de histogramas, como pode ser visto nas Figuras 4 e 5, a partir de intersecções entre histogramas para amenizar problemas como distrações no plano de fundo, visão de um objeto em diferentes pontos de vista, oclusão, variação de resolução e condições de iluminação da imagem (SWAIN; BALLARD, 1991).

Um ponto a se observar é que o histograma gerado irá conter o número de *pixels* da imagem na totalidade, ou seja, se a imagem possui uma escala de 400 de largura por 400 de altura ela possui um total de 160.000 *pixels* que serão contabilizados e darão origem ao histograma, logo quanto maior a resolução de uma imagem mais complexo será seu histograma.

- **Descritores Baseados em forma ou contorno**

As formas são características extremamente importantes ao longo da história humana. Em estudos médicos, as técnicas de extração de formas têm se destacado para auxiliar no diagnóstico, muitas vezes combinadas com outros descritores para obter maior precisão (ZHANG et al., 2022; PEDRONETTE; TORRES, 2010; AFIFI; ASHOUR, 2012). Descritores baseados em forma necessitam da aplicação de um pré-processamento para identificação e extração de contornos, por meio de um algoritmo de detecção de bordas, como o de Canny, Sobel ou Prewit, por exemplo. A borda de um objeto em uma imagem é identificada mediante uma súbita mudança de contraste ou cor. Com o contorno extraído, a imagem pode ser dividida em grupos de *pixels* para a contagem de linhas e construção de um histograma.

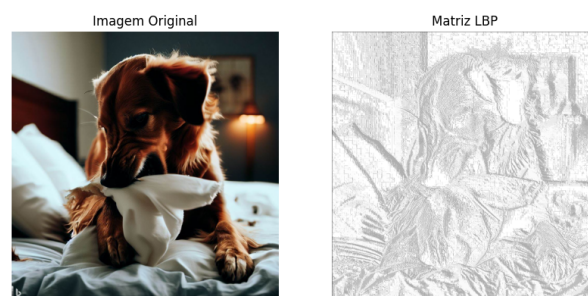
Como exemplo dos descritores baseados em contorno podemos citar o *Contour Features Descriptor* (CFD) o conceito básico das técnicas de extração de características é utilizar funções que calculam o valor das características para cada ponto do

contorno do objeto, analisando os *pixels* vizinhos. Esse processo é faseado, sendo suas fases: segmentação de imagem, extração de contornos, normalização de contornos, extração das características, construção do descritor de características e comparação dos descritores (PEDRONETTE; TORRES, 2010). É importante destacar que o descritor pode variar conforme a implementação e as necessidades específicas do problema em questão.

- **Descritores baseados em textura**

Descritores baseados em textura são características quantitativas extraídas de uma imagem para capturar e representar padrões e estruturas texturais presentes nela. Esses descritores buscam quantificar informações sobre variações locais de intensidade, contraste, rugosidade e outros atributos visuais que definem a textura de uma região na imagem (VALEM, 2019). Entre os descritores de texturas pode-se utilizar como exemplo o *Local Binary patterns* (LBP) que se trata de um operador simples e poderoso que rotula os *pixels* de uma imagem a partir de um *trasholding* de vizinhança de cada *pixel* e considera o resultado como um número binário. Devido a sua simplicidade e poder de descrição, essa abordagem se tornou extremamente popular, tendo como possível propriedade na aplicação no mundo real sua robustez a mudanças causadas por imagens em tons de cinza, como variações de iluminação (PIETIKÄINEN, 2010; KAUR; NAZIR et al., 2021), ilustrado pela Figura 6.

Figura 6 – Ilustração de uma matriz LBP.



Fonte: Adaptado de (MICROSOFT, 2023).

- **Descritor baseado em gradiente**

O *Histograma de Gradientes Orientados* (HoG) é um método que segmenta a imagem em conjuntos de *pixels* e descreve a imagem usando vetores direcionados, dos quais um histograma é calculado. Esse descritor têm sido extensivamente estudado na literatura, especialmente para reconhecimento facial, além de ser objeto de estudos aplicados

em diagnósticos médicos (ANDREA; IKHSAN; SUDIRMAN, 2022; AYALEW et al., 2022; BAO et al., 2020; ENG et al., 2019). A Figura 7 ilustra esse método.

Figura 7 – Ilustração de uma matriz HoG.



Fonte: Adaptado de (MICROSOFT, 2023).

- **Descritores locais**

Um dos maiores problemas enfrentado pelos descritores globais se encontra no chamado problema de variância, que se trata da dificuldade de identificação em alguns casos onde o objeto se encontra rotacionado ou transladado. Para suprir o problema de variância e ampliar a eficácia das buscas em sistemas baseados em CBIR foram propostos os descritores locais, os quais geralmente utilizam abordagens de *Bag of Visual Words* (BoVW) (VALEM, 2019).

A aplicação das técnicas de BoVW são baseadas nas abordagens de *Bag of Words* (BoW). Abordagens baseadas em BoVM analisam diferentes imagens em buscas de *Keypoints* que possuam relevância para descrição da imagem e sejam invariantes. As partes relevantes são utilizadas para criação de um dicionário de palavras visuais, podendo ser criado um vetor de características a partir da ocorrência de cada elemento contido no dicionário criado (AWAD; HASSABALLAH, 2019; ABOUZAHIR; SADIK; SABIR, 2021).

Como exemplos de descritor local podemos destacar o *Oriented FAST Rotate BRIEF* (ORB), que realiza a extração de *keypoints*, emprega o detector chamado *Features from Accelerated Segment Test* (FAST) e adota a abordagem *Binary Robust Independent Elementary Features* (BRIEF) para a extração das descrições das características da imagem. Utilizando o método de detecção de cantos de Harris (*Harris Corner Detector*), o ponto de interesse mais apropriado é identificado dentre aqueles detectados pelo método FAST. Para orientar os *keypoints*, é aplicada uma orientação nos cantos e uma rotação é executada no teste de padrões binários, utilizando a direção de correção. O vetor resultante de descrição para um *keypoint* é um vetor de características com 32 dimensões (RUBLEE et al., 2011).

- **Descritores Baseados em Redes Neurais Convolucionais**

São modelos computacionais inspirados no funcionamento do cérebro humano. Eles passam por treinamento em grandes datasets de imagens para aprender a extrair características relevantes, como texturas, formas, cores e outros padrões visuais. Esses descritores podem mapear automaticamente as informações visuais de uma imagem para um espaço de características, onde cada imagem é representada por um vetor numérico. Redes de aprendizado profundo são geralmente treinadas de maneira supervisionada e seus vetores de peso podem ser utilizados para extrair características de uma imagem. Um tipo de rede *deep learning* comumente utilizada são as redes convolucionais (CNN) (VALEM; PEDRONETTE, 2019; ZHANG et al., 2022).

Composta por nós interconectados, chamados de neurônios, que trabalham juntos para processar e aprender a partir dos dados de entrada. Cada neurônio recebe sinais de entrada, realiza uma operação matemática neles e produz uma saída. Essas saídas são então passadas para outros neurônios, criando uma rede de fluxo de informações. A força das conexões entre os neurônios, conhecida como pesos, determina a importância de cada sinal de entrada na computação geral (STUART; PETER, 2010; KOVÁCS, 2002).

Uma rede neural convolucional (CNN) é um tipo especializado de rede neural utilizado principalmente para o processamento de imagens. Ela foi projetada para reconhecer padrões e características específicas nas imagens. Uma característica chave das CNNs é a camada convolucional, que consiste em filtros de convolução. Esses filtros são pequenas janelas deslizantes que percorrem a imagem, capturando informações locais. A convolução é uma operação matemática que combina os valores dos *pixels* da imagem com os pesos do filtro, produzindo um mapa de características. À medida que a informação é processada nas camadas convolucionais, a rede neural consegue detectar características de níveis cada vez mais altos. Além das camadas convolucionais, as CNNs também podem incluir camadas de *pooling*, que reduzem a dimensionalidade dos mapas de características, e camadas totalmente conectadas, que realizam o ranqueamento final. Essas camadas adicionais ajudam a extrair informações relevantes e a reduzir a quantidade de parâmetros necessários para o treinamento (LI et al., 2021b).

Vale ressaltar que as redes neurais enfrentam um desafio conhecido como exemplo adversário, que envolve a inclusão proposital de interferências durante o treinamento da rede. Esse desafio pode ser dividido em dois tipos: contaminação (*poisoning*) e evasão (*evasion*). No ataque de contaminação, a interferência é direcionada para contaminar os dados de treinamento durante o processo de treinamento do modelo. Um modo convencional de contaminação é o ataque de *backdoor*, onde um padrão

é inserido nos dados de treinamento para influenciar os resultados futuros. Já nos ataques de evasão, as interferências ocorrem geralmente após o treinamento do modelo. Nesses ataques, um padrão é adicionado aos dados de entrada para produzir resultados completamente incorretos ou indesejáveis (ALDAHDOOH et al., 2022).

Como exemplo de descritores baseados em redes neurais convolucionais ou de aprendizado profundo pode-se destacar:

– ***Dual Path Networks (DPN)***

Redes Dual Path Networks (DPNs) são arquiteturas avançadas de redes neurais profundas que possuem duas vias paralelas e independentes de processamento de informações, que visam aprimorar a eficiência de aprendizado de representações. Elas se destacam por sua capacidade de capturar diferentes aspectos dos dados e alcançar maior robustez e precisão em tarefas de visão computacional, como classificação e detecção de objetos. Através do caminho ResNet utiliza blocos residuais, conhecidos por sua habilidade de minimizar o problema de degradação nas redes profundas, permitindo que a rede aprenda mapeamentos de identidade. Essa via foca em captar informações globais, como estruturas e bordas de alta e média frequência, essenciais para o reconhecimento de padrões principais nas imagens, enquanto no caminho DenseNet a estrutura densa conecta cada camada a todas as camadas subsequentes, promovendo a reutilização de características e melhorando o fluxo de informações. Essa via ajuda a capturar detalhes mais sutis e combinações de características, enriquecendo a representação e a precisão da rede. (CHEN et al., 2017)

– ***Residual Network (RESNET)***

O conceito central de uma ResNet é o bloco residual, pois em uma rede neural comum, cada camada aprende a transformar diretamente a entrada em uma saída desejada. No entanto, com o aumento da profundidade, torna-se mais difícil para cada camada aprender com eficiência, pois as camadas mais profundas têm dificuldade em captar os gradientes dos erros iniciais. A ResNet resolve isso com o uso de uma estrutura de atalho, conhecida *skip connection*, dentro do bloco residual. O bloco residual consiste em duas ou três camadas com uma *skip connection*, que pula uma ou mais camadas, levando a entrada diretamente para a saída do bloco, representado matematicamente como $H(x)$ como a função de mapeamento desejada para uma camada ou bloco da rede. No entanto, a ResNet introduz uma função residual $F(x)$, definida pela diferença entre a função de mapeamento ($H(x)$) e a entrada x , expressa por $F(x) = H(x) - x$. Assim, o bloco residual aprende essa função residual e a saída do bloco se torna $S_{F(x)} = F(x) + x$ (HE et al., 2016)

– ***Squeeze-and-Excitation Network (SENET)***

A *Squeeze-and-Excitation Network* (SENet) é uma arquitetura de rede neural projetada para melhorar a capacidade da rede de extrair características relevantes em imagens, aumentando a eficiência do aprendizado através da introdução do mecanismo de compressão e excitação (*Squeeze-and-Excitation*) responsável por adaptar dinamicamente a importância de cada dimensão de profundidade das características extraídas.

Se baseia no bloco de compressão e excitação, que consiste em três fases principais: *Squeeze* - através da aplicação de uma média global em cada canal de saída convolucional, exerce uma compressão reduzindo as dimensões espaciais para um único valor; *Excitation* - a rede passa o vetor comprimido por uma camada de ativação, comunmente composta por duas camadas totalmente conectadas, para capturar interdependências não-lineares entre os canais, expandindo a dimensionalidade para o valor original de canais; *Reweighting* - responsável por multiplicar o canal pelo seu valor de ponderação aprendido, ajustando dinamicamente a importância de cada canal antes de passar a saída para a próxima camada da rede (HU; SHEN; SUN, 2018).

– ***Extreme Inception (XCEPTION)***

A *Extreme Inception Network* (Xception) é uma arquitetura de rede neural profunda que expande os conceitos da rede Inception, focando em eficiência e performance. Ela é baseada em uma versão extrema e simplificada do módulo Inception, usando uma combinação de convoluções separáveis em profundidade e convoluções ponto a ponto para capturar características complexas em imagens de forma mais eficaz e com menor custo computacional.

Baseia-se no uso dos princípios de blocos residuais, como visto na rede DPN, e convoluções separáveis em profundidade, composta pelas etapas de convolução em profundidade que aplica uma convolução para cada canal de entrada individualmente, e a convolução ponto a ponto que combina os canais por meio de uma convolução de 1×1 , captando interdependências entre eles.

Cada bloco Xception consiste em uma sequência de convoluções separáveis em profundidade e convoluções ponto a ponto, seguidas de uma camada de ativação, estrutura essa que se repete em vários blocos, alternando operações para capturar relações espaciais e dependências entre os canais (CHOLLET, 2017).

• **Descritores Baseados em Redes Neurais *Transformers***

As redes neurais *Transformers* têm sido objeto de considerável atenção na área de visão computacional devido à sua capacidade conhecida de lidar com tarefas complexas, como o reconhecimento de expressões faciais. Essas redes apresentam uma complexidade temporal proporcional a $O(n^2)$, onde n é o comprimento da

sequência de entrada. Para mitigar essa complexidade temporal, destacam-se técnicas como a atenção esparsa baseada em localização e a autoatenção (TEJASHWINI; ARADHANA, 2023).

Dentre os tipos de *transformers*, destacam-se: modelos de CNN aprimorados pela aplicação de redes *Transformers* visuais, capazes de adicionar uma camada de autoatenção ou substituir convoluções por blocos de autoatenção; *Transformers* hierárquicos e profundos, onde os hierárquicos substituem a estrutura colunar de resolução fixa por uma configuração piramidal, enquanto os *Transformers* profundos evitam a suavização excessiva; *Transformers* visuais, arquiteturas aplicadas à área de visão computacional para tarefas de classificação, detecção e segmentação; *Transformers* autosupervisionados, modelos dominantes em tarefas de processamento de linguagem natural; e *Transformers* visuais autosupervisionados, utilizados em tarefas de segmentação, rastreamento e geração de imagens. Na área de aplicação para extração de características, as redes *Transformers* demonstram grande eficácia, superando o desempenho das redes convolucionais tradicionais (LIU et al., 2023a)

Como exemplos de descritores baseados em Redes transformadoras, pode-se destacar:

– ***Swin Transformer (SWINTF)***

É uma arquitetura de rede neural baseada em Transformers, desenvolvida especificamente para tarefas de visão computacional, como classificação de imagens, detecção e segmentação. Ela utiliza um mecanismo de atenção hierárquica em janelas deslizantes, o que melhora a eficiência computacional e permite lidar com imagens de alta resolução.

Seu funcionamento se baseia em cinco estruturas: divisão de *patches* - responsável por dividir a imagem em pequenas regiões (*patches*), onde cada *patch* é tratado como um *token* que representa uma parte da imagem; mecanismo de atenção em janelas - aplica a atenção apenas dentro de janelas menores, o que reduz o número de operações e aumenta a eficiência; janelas deslizantes com mudança de janelas - visando capturar informações contextuais além de uma única janela, o SWINTF utiliza-se do mecanismo de janelas deslizantes, onde em uma etapa as janelas estão em uma posição fixa, e na próxima, são deslocadas uma quantidade determinada de *pixels*; arquitetura hierárquica - utiliza-se de uma estrutura hierárquica, onde a resolução dos patches é gradualmente reduzida à medida que a profundidade da rede aumenta. Isso é feito por meio de camadas de poolings, que ajudam a extrair características em várias escalas e a consolidar a informação; camadas de perceptron multicamada e normalização - Após cada bloco de atenção, uma camada de perceptron multicamada com normalização de camada é aplicada para ajustar as representações e promover a estabilidade no aprendizado (LIU et al., 2022).

– **Visual Transformer (ViT)**

É uma arquitetura de rede neural que adapta o modelo de Transformers, originalmente desenvolvido para Processamento de Linguagem Natural, para tarefas de visão computacional, como classificação de imagens. Em vez de usar convoluções, o ViT modela a imagem como uma sequência de tokens e aplica o mecanismo de atenção global para aprender relações entre partes da imagem. Inicialmente divide a entrada em pequenos blocos (*patches*) de tamanho fixo, como 16×16 ou 32×32 , e achata cada bloco em um vetor unidimensional. Para solucionar a falta de um conceito de ordem espacial, utilizam-se codificações de posição para cada *patch* na imagem, que adicionam informações sobre a posição de cada patch na imagem, garantindo que o modelo capture a localização relativa dos elementos visuais.

Após a criação dos tokens, o ViT aplica o mecanismo de atenção global a esses tokens. Cada camada de atenção calcula a relação entre todos os pares de patches (tokens) da imagem, permitindo que o modelo entenda como cada parte da imagem se relaciona com as demais. Adiciona um token especial de classificação no início da sequência de tokens. Esse token, que é processado juntamente com os patches, captura uma representação global da imagem ao final do processo de atenção e é usado para a tarefa de classificação. Cada camada do Transformer no ViT utiliza múltiplas cabeças de atenção, permitindo que o modelo foque em diferentes aspectos da imagem simultaneamente.

Possui diferentes tamanhos de arquitetura que se diferenciam pela quantidade de blocos de atenção, cabeças de atenção e dimensão do *embedding*, sendo esses tamanhos: *Base* (B) com 12 blocos de atenção, 12 cabeças de atenção e dimensão do *embedding* igual à 768; *Large* (L) com 24 blocos de atenção, 16 cabeças de atenção e dimensão do *embedding* igual à 1024; *Huge* (H) com 32 blocos de atenção, 16 cabeças de atenção e dimensão do *embedding* igual à 1280. Juntando o tamanho dos patches e o tamanho de arquitetura obtemos a denominação do modelo ViT utilizado, como exemplo ViT-B16 ou ViT-L32 (LIU et al., 2023b).

– **Distillation with No Labels Version 2 (DINOv2)**

Trata-se de uma arquitetura que aprimora o aprendizado auto-supervisionado, permitindo que modelos de visão computacional aprendam representações significativas sem precisar de rótulos, ou seja, sem supervisão explícita de classes. Desenvolvida especialmente para lidar com grandes conjuntos de dados de imagens, essa abordagem usa a destilação de conhecimento sem rótulos para treinar redes profundas, transformando o modelo em um método poderoso para tarefas de visão, como classificação, detecção e segmentação de imagens, mesmo sem dados rotulados.

O funcionamento da DINOv2 se baseia em um processo de destilação auto-supervisionado que envolve dois papéis principais: o modelo professor - modelo estável cujos pesos são atualizados de forma mais lenta, usando geralmente a técnica de atualização exponencial média (momentum), gerando representações-alvo para o modelo aluno; O modelo aluno - treinado para imitar o modelo professor, ajustando seus pesos para gerar representações semelhantes para cada vista de uma mesma imagem.

Esse modelo pode ser utilizado com uma base no modelo ViT, gerando modelos como DINOv2 ViT-B16, por exemplo, e apresenta um bom equilíbrio entre precisão e custo (OQUAB et al., 2023).

2.2 Métricas de Avaliação de Resultados

Para avaliar sistemas de Recuperação da informação (RI), sejam para recuperação de textos ou imagens, é necessário a aplicação de uma ou mais das denominadas métricas de avaliação de resultados, as quais são cálculos que se baseiam na avaliação do resultado obtido comparado à coleção ao qual o algoritmo de RI foi aplicado (BAEZA-YATES; RIBEIRO-NETO, 2013).

Dentre as métricas de avaliação pode-se destacar:

- **Precisão (*Precision*) e Revocação (*Recall*)**

Para ambos considera-se uma consulta I em uma coleção de referência e o conjunto R os documentos relevantes, sendo $|R|$ o número de documentos nesse conjunto. Supondo que o algoritmo que será avaliado processa I e gera um conjunto resposta A , sendo $|A|$ a composição nesse conjunto e $|R \cap A|$ a intersecção de elementos dos conjuntos A e R (FRÄNTI; MARIESCU-ISTODOR, 2023). Considerando os dados descritos, pode-se definir a **precisão** como a divisão da intersecção dos elementos dos conjuntos A e R pelos elementos do conjunto de documentos recuperados $|A|$, visto na Equação 2.1:

$$P(r) = \frac{|R \cap A|}{|A|} \quad (2.1)$$

No caso da **revocação** a intersecção dos elementos dos conjuntos A e R é dividida pelo conjunto de elementos relevantes $|R|$, com visto na Equação 2.2:

$$R(r) = \frac{|R \cap A|}{|R|} \quad (2.2)$$

- **Mean Average Precision (*MAP*)**

A abordagem MAP têm como conceito gerar um valor de média do ranking de precisão obtido após cada documento relevante. Considerando N_q como o número total de consultas, R o conjunto de documentos relevantes, sendo $|R|$ o número de documentos relevantes para a consulta q_i (BAEZA-YATES; RIBEIRO-NETO, 2013). Tomando $R_i[k]$ como sendo o k -ésimo documento de R_i , logo $P(R_i[k])$ trata-se da precisão de R_i no *ranking* de q_i , quando não recuperado nenhum documento o valor retornado é igual a zero, logo a expressão matemática do MAP equivale à:

$$MAP = \frac{1}{N_q} \sum_{i=1}^{N_q} \frac{1}{|R_i|} \sum_{k=1}^{|R_i|} P(R_i[k]) \quad (2.3)$$

- ***E-Score***

O *E-score* permite a combinação dos valores de precisão e revocação ao mesmo tempo que permite a especificação de peso entre eles, onde $r(j)$ se trata da revocação da posição j e $P(j)$ a precisão da posição j no ranking, $b \geq 0$ é o parâmetro de peso que o usuário irá definir, para valores de $b \geq 1$ o usuário apresenta mais interesse na revocação, quando b é igualado a 1 temos a medida do *F-Score* também chamada de média harmônica, lembrando que o resultado obtido estará no intervalo $0 \leq E(j) \leq 1$ (BAEZA-YATES; RIBEIRO-NETO, 2013). A equação do *E-Score* é dada por:

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}} \quad (2.4)$$

- ***F-Score***

O *F-score* é um complemento a medida do *E-Score*, retornando 1 quando todos os documentos são relevantes e 0 quando nenhum documento for relevante, na média harmônica utiliza-se a equação (FRÄNTI; MARIESCU-ISTODOR, 2023):

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}} \quad (2.5)$$

3 Abordagens de Combinação Não Supervisionadas

De forma geral, o conceito de combinação têm como alvo a fusão de resultados obtidos por algoritmos distintos, para uma consulta efetuada. Abordagens não supervisionadas não necessitam da intervenção do usuário, dados rotulados ou treinamento (KUNCHEVA, 2014).

3.1 Definição Formal

Neste trabalho, conforme a definição formal apresentada por Pedronette, Penatti e Torres (2014), será empregada a seguinte notação ao longo deste capítulo: $\mathcal{C} = \{img_1, img_2, \dots, img_n\}$ representa uma coleção de imagens, onde n é o tamanho da coleção. $\mathcal{R} = \{\tau_1, \tau_2, \dots, \tau_n\}$ sendo o conjunto de listas ranqueadas de cada imagem da coleção $\mathcal{C}$ e $\mathcal{D}$ um descritor que define uma função de distância entre duas imagens img_i e img_j como $\rho(img_i, img_j)$, simplificado como $\rho(i, j)$.

A partir da função de distância ρ , é possível calcular uma lista ranqueada τ_q com base na imagem de consulta img_q . Essa lista ranqueada contém os resultados da medida de similaridade para toda a coleção, e espera-se que as melhores posições da lista contenham as imagens mais relevantes para a consulta. A lista $\tau_q = (img_1, img_2, \dots, img_{n_s})$ representa a permutação do subconjunto $C_s \subset C$, que contém as imagens mais similares à imagem de consulta img_q , onde t equivale aos elementos de C_s , $|C_s| = t$. Denota-se por $t_q(i)$ uma permutação da função bijectiva do conjunto C_s para o conjunto $[n_s] = 1, 2, \dots, n_s$.

Em alguns momentos pode ser realmente desejado considerar apenas um subconjunto n_s de imagens mais similares em τ_q , onde $n_s \ll n$ é o número de imagens nas primeiras posições da lista. Essa estratégia é extremamente útil principalmente para grandes coleções, pois com uma quantidade de imagens muito grande o processo para calcular a lista resultante se torna altamente custoso.

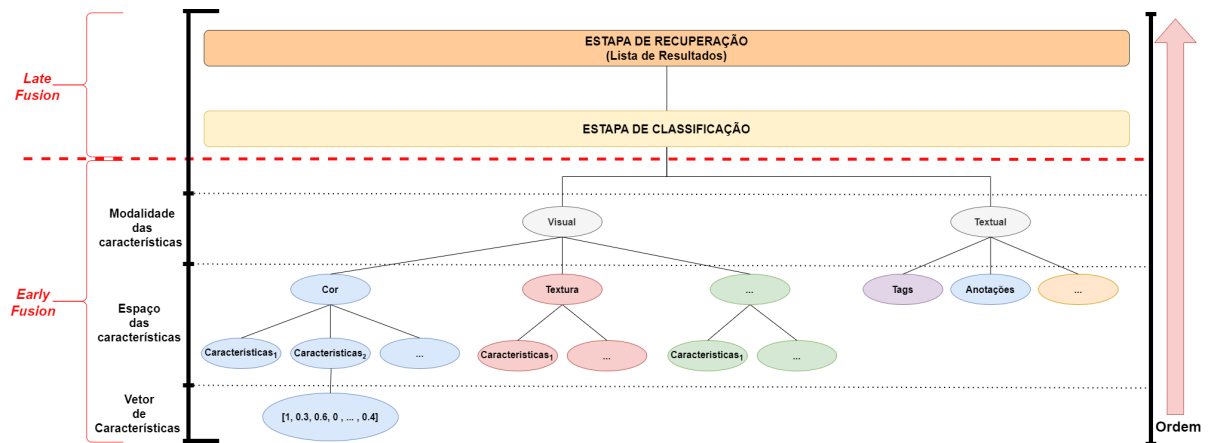
Assim pode-se dizer que $\mathcal{R}$ equivale a $\tau_{conj} = \{\tau_1, \tau_2, \dots, \tau_m\}$ sendo o conjunto de rankings a serem combinados, cujo tamanho total do subconjunto equivale a m , buscando como resultado uma lista ótima τ' .

3.2 Técnicas de Agregação

Técnicas de combinação, ou também chamadas de técnicas de fusão, possuem uma divisão de abordagens em duas classes bem definidas, *Early Fusion* e *Late Fusion*. A

Figura 8 ilustra a etapa onde cada classificação de fusão ocorre. É possível encontrar na literatura, em alguns casos, a citação de uma terceira divisão denominada *Intermediate Fusion*, que consiste no treinamento efetuado nos classificadores a fim de obter a melhor combinação entre os vetores de entrada, assim o treinamento afeta o vetor de características gerando uma nova lista de resultados (IONESCU et al., 2014).

Figura 8 – Ilustração da esquemática de um sistema baseado em CBIR.

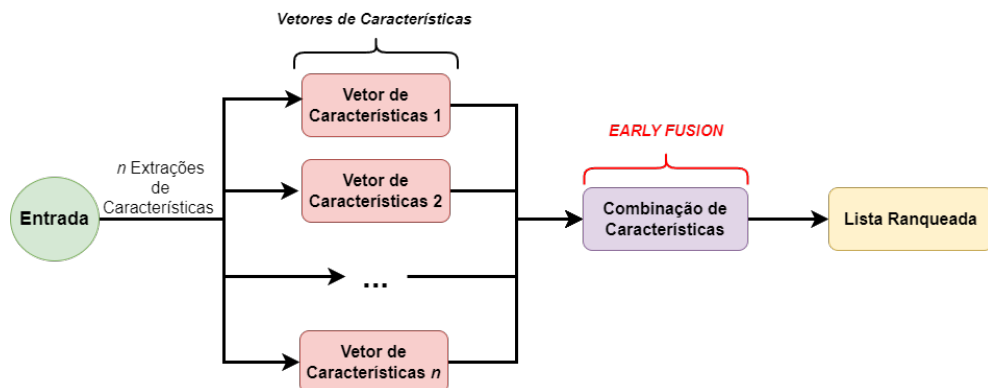


Fonte: Adaptado de (PIRAS; GIACINTO, 2017).

• *Early Fusion*

Na combinação prematura (*early fusion*) a combinação das características ocorre com o resultado dos vetores de características unimodais, ou seja, antes da aplicação do cálculo das distâncias, como ilustrado pela Figura 9, resultando em uma única lista de resultados unificada (SNOEK; WORRING; SMEULDERS, 2005).

Figura 9 – Ilustração do fluxo de combinação de um sistema baseado em *Early Fusion*.

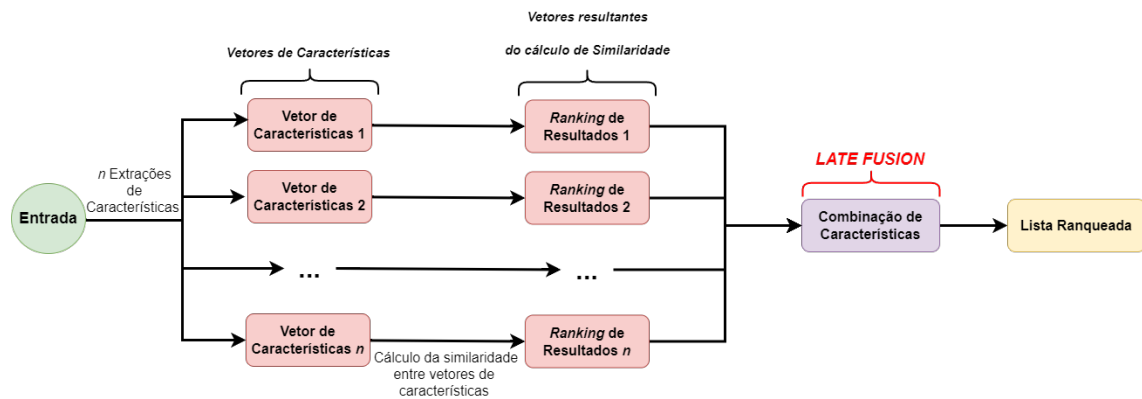


Fonte: Adaptado de (SNOEK; WORRING; SMEULDERS, 2005).

• *Late Fusion*

Na combinação tardia (*late fusion*) a combinação das características ocorrem somente após o cálculo das distâncias de cada vetor de características, como ilustrado pelo Figura 10, assim aplicando a metodologia de agregação a partir do vetor de similaridade e não do vetor de características diretamente (YAO et al., 2021).

Figura 10 – Ilustração do fluxo de combinação de um sistema baseado em *Late Fusion*.



Fonte: Adaptado de (SNOEK; WORRING; SMEULDERS, 2005).

Dentre os principais métodos de combinação pode-se destacar combinação aritmética não treinável, combinação conservativa, agregação de listas ranqueadas, combinações de núcleos, abordagens estatísticas e combinação baseada em programação genética (VALEM, 2019). Este trabalho tem como foco as técnicas de agregação de listas ranqueadas, as quais serão tratadas na Seção 3.3 sendo classificadas como *Late Fusion*.

3.3 Técnicas de Agregação de Listas Ranqueadas

As técnicas de agregação de listas ranqueadas visam combinar as listas ranqueadas obtidas por meio dos descritores mencionados na Seção 2.1, a fim de melhorar os resultados obtidos. Essa abordagem visa aproveitar os elementos que são bem ranqueados em múltiplas listas, pois esses têm maior probabilidade de atender com maior precisão à consulta realizada. Quanto mais vetores de características forem comparados, mais preciso tende a ser o resultado (SCHALEKAMP; ZUYLEN, 2009; BAŁCHANOWSKI; BORYCZKA, 2022).

Dentre as principais técnicas de agregação de listas ranqueadas destacam-se:

- **Borda Count:**

A contagem de Borda (*Borda count*) visa unificar o resultado de decisões multi-agentes (listas ranqueadas geradas por algoritmos distintos), onde todas alternativas (listas ranqueadas resultantes) devem ser consideradas (GARCÍA-LAPRESTA;

MARTÍNEZ-PANERO; MENESES, 2009). A aplicação desta abordagem resulta em uma lista ótima, a partir da ordenação crescente de cada elemento, com a aplicação matemática da Equação 3.1 a fim de definir uma pontuação para cada elemento da coleção:

$$Bc_i = \sum_{j=0}^t \tau_j(i) \quad (3.1)$$

- **Métodos família COMB**

Os métodos pertencentes à família COMB são técnicas de fusão empregadas para determinar a pontuação de imagens ou documentos conforme seu ranqueamento. Na Tabela 2, são explicitados os métodos COMB e a partir da ordenação crescente dos elementos, obtém-se uma lista τ' .

Os métodos lineares de combinação executam a agregação das listas ranqueadas a partir do uso de uma função de atribuição de pontos para todo item i da coleção a partir de sua posição nas listas ranqueadas τ_j (SHAW; FOX, 1995), conforme a Equação 3.2:

$$S_C(i, \tau_j) = 1 - \frac{\tau_j(i) - 1}{|\tau_j|} \quad (3.2)$$

A normalização descrita pode ser aplicada aos métodos COMB, apresentados na Tabela 2, e a partir da ordenação crescente de cada elemento, pode ser uma lista resultante τ' :

Tabela 2 – Métodos lineares de combinação (COMB).

Algoritmo	Equação
CombMIN	$\min_{\tau_j \in \tau_{conj}} S_C(i, \tau_j)$
CombMAX	$\max_{\tau_j \in \tau_{conj}} S_C(i, \tau_j)$
CombSUM	$\sum_{\tau_j \in \tau_{conj}} S_C(i, \tau_j)$
CombMED	$\frac{1}{ \tau_{conj} } \times CombSUM(i, \tau_{conj})$
CombANZ	$\frac{1}{ \tau_i \tau_i \in \tau_{conj} : i \in \tau_i } \times CombSUM(i, \tau_{conj})$
CombMNZ	$\{ \tau_i \tau_i \in \tau_{conj} : i \in \tau_i \} \times CombSUM(i, \tau_j)$

Vale ressaltar que tanto na equação CombANZ quanto na CombMNZ a análise ignora as k primeiras posições ao calcularem o valor de $|\tau_i| \tau_i \in \tau_{conj} : i \in \tau_i|$ (VALEM, 2019; RENDA; STRACCIA, 2003);

- **Reciprocal Rank Fusion**

Calcula uma pontuação para cada elemento, porém considera apenas os elementos presentes nas k primeiras posições, como visto na Equação 3.3, a fim de mitigar o efeito de resultados de ranqueamento muito baixos, ou seja, maiores que k (*outlier*), e através da ordenação crescente se obtém uma nova lista τ' (CORMACK; CLARKE; BUETTCHER, 2009), expressa pela Equação 3.3 :

$$Rf_i = \sum_{j=0}^t \frac{1}{k + \tau_j(i)} \quad (3.3)$$

- ***Median Rank Aggregation (MRA)***

Através da aplicação do cálculo de mediana das listas ranqueadas τ para cada elemento, i uma nova lista τ' é criada (FAGIN; KUMAR; SIVAKUMAR, 2003b).

Na literatura, é possível encontrar a aplicação do MRA (Método de Regressão Agregada) em conjunto com a distância de Kemeny para a resolução do problema da ordem ideal de baldes (*optimal bucket order problem*). O objetivo desse problema é determinar a ordem mais eficiente para colocar os itens nos baldes, de modo a minimizar ou maximizar algum critério específico. Esse critério pode variar dependendo do contexto do problema, como otimizar o tempo de acesso aos itens, maximizar o uso dos recursos dos baldes ou minimizar o espaço desperdiçado (D'AMBROSIO et al., 2019; GIONIS et al., 2006).

- ***Algoritmo de Aprendizado Não Supervisionado para Agregação de Listas Ranqueadas (ULARA)***

O algoritmo de aprendizado não supervisionado para agregação de listas ranqueadas ULARA (*Unsupervised Learning Algorithm for Rank Aggregation*) é um método específico que combina conceitos de aprendizado não supervisionado com a tarefa de agregação de listas ranqueadas, abordando o desafio de combinar várias listas ranqueadas sem ter rótulos ou supervisão externa para orientar o processo. Ele é usado quando diferentes listas ranqueadas ou fontes fornecem suas próprias listas ranqueadas para um conjunto de itens em comum, e é necessário encontrar uma lista agregada.

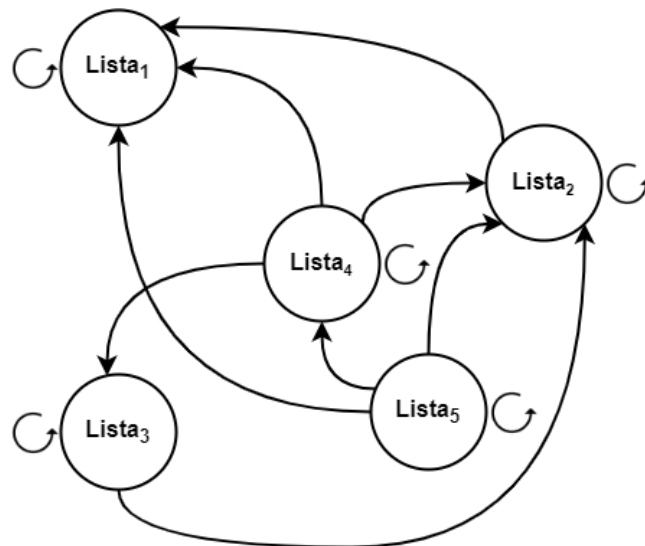
Esse algoritmo de aprendizado não supervisionado para agregação de listas ranqueadas ULARA é útil em várias aplicações que envolvem a combinação de listas ranqueadas de diferentes fontes, como sistemas de recomendação, avaliação de produtos e tomada de decisões coletivas. O ULARA pode considerar características adicionais dos itens, bem como a confiabilidade das fontes, para aprimorar o ranqueamento agregada. O algoritmo continua iterando até que alcance uma convergência e o ranqueamento agregado seja considerado estável (KLEMENTIEV; ROTH; SMALL, 2007).

- **Cadeia de Markov**

É um método que combina os conceitos de agregação de listas ranqueadas e cadeia de Markov para obter um ranqueamento agregado a partir de múltiplas listas ranqueadas. É útil quando se deseja combinar várias listas ranqueadas de diferentes fontes coerentemente. Ele considera as preferências coletivas e a estrutura das transições entre os itens para obter um ranqueamento agregado.

O processo começa com um conjunto de listas ranqueadas, onde a ideia é construir uma cadeia de Markov na qual os estados correspondam aos itens a serem ranqueados e as transições entre os estados sejam governadas pelas probabilidades de transição entre os itens nas listas. Uma vez que a cadeia de Markov tenha sido definida, é possível aplicar técnicas de análise da cadeia de Markov para obter a agregação das listas ranqueadas. Isso é geralmente feito por meio do cálculo da distribuição estacionária da cadeia de Markov, que indica a probabilidade de estar em cada estado após um número suficientemente grande de transições. A distribuição estacionária pode ser interpretada como um ranqueamento agregado, no qual os itens com maiores probabilidades de serem visitados representam um ranqueamento mais alta (SCULLEY, 2007).

Figura 11 – Agregação com cadeia de Markov, executando a agregação das listas.



Fonte: Adaptado de (SCULLEY, 2007)

Na Figura 11 é ilustrado o processo de agregação através da cadeia de Markov, onde o número de possíveis transições é determinado pela escolha do método de mapeamento, enquanto as auto-transições representam a probabilidade de se manter naquele nó durante um passo.

3.4 Técnicas de Seleção de Descritores Não Supervisionada

A seleção apropriada de descritores desempenha um papel crucial na obtenção de resultados precisos e relevantes na recuperação de imagens baseadas em conteúdo. A seleção dos descritores influencia diretamente a composição da lista resultante, determinando a qualidade das informações recuperadas. Esta etapa é fundamental, uma vez que os descritores atuam como representações sintetizadas das características ou propriedades dos dados, e a seleção inadequada pode comprometer significativamente a acurácia e a utilidade dos resultados obtidos. Nesta seção, será explorada a importância crítica da seleção criteriosa de descritores, destacando seu impacto na precisão, relevância e eficácia da lista final de elementos analisados, divididas entre estimativas de eficácia na Seção 3.4.1 e métricas de correlação na Seção 3.4.2, que apesar de não se tratarem de métricas de seleção, são utilizadas para avaliar a diversidade entre os descritores e assim considerar aspectos de complementariedade.

3.4.1 Estimativas de Eficácia

Estimativas de eficácia são fundamentais e procuram deduzir a eficácia de uma lista ranqueada sem a necessidade de uma rotulagem dos dados, além de poderem ser utilizadas para uma seleção das combinações.

Essas estimativas exploram o conceito de hipótese de agrupamentos (*cluster Hypothesis*), que calculam os valores entre $[0, 1]$ e quanto mais próximo de 1 melhor o resultado. Então, a hipótese de agrupamentos defende que itens similares possuem alta semelhança e assim tendem a possuir elementos em comum nas k primeiras posições (JARDINE; RIJSBERGEN, 1971; KURLAND, 2014; CRESTANI; WU, 2006).

Dentre as medidas de estimativa de eficácia não supervisionadas se destacam:

- ***Authority Measure***

A métrica da medida de autoridade estima o quão bem uma lista ranqueada reflete a ideia de que imagens similares devem aparecer frequentemente no topo das listas ranqueadas umas das outras (Hipótese de Cluster). É denotada como $A_s(q, k)$, e representa uma estimativa da eficácia com base na densidade do grafo de vizinhança em uma lista classificada de uma imagem q . Aqui, $\mathcal{N}(q, k)$ denota o conjunto dos primeiros k elementos mais próximos de q na lista τ_q .

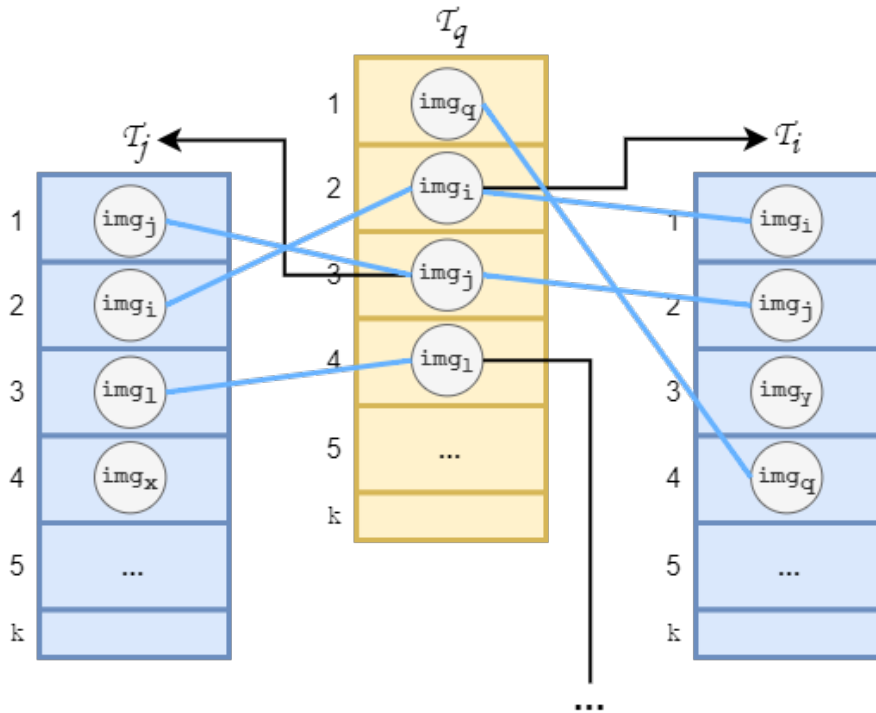
A pontuação é calculada analisando a densidade de referências recíprocas entre as k posições superiores de uma lista ranqueada. Especificamente, ela soma as pontuações de autoridade ao quadrado de todas as listas ranqueadas onde tanto img_q quanto img_i aparecem entre as k posições superiores. Uma pontuação mais alta indica um subgrafo mais denso de referências recíprocas, sugerindo uma lista ranqueada mais

confiável e eficaz. Matematicamente, essa métrica é formalizada pela Equação 3.4 para o intervalo $[0, 1]$, com 1 representando um grafo de k vizinhos mais próximos (kNN) recíproco perfeitamente conectado, onde todas as k imagens superiores se referenciam umas às outras. Isso significa que o algoritmo considera as k imagens superiores na lista ranqueada como altamente relacionadas e confiáveis para análise posterior (PEDRONETTE; PENATTI; TORRES, 2014).

$$A_s(q, k) = \frac{\sum_{i \in \mathcal{N}(q, k)} \sum_{j \in \mathcal{N}(i, k)} f_{in}(j, q)}{k^2} \quad (3.4)$$

A Figura 12 ilustra o funcionamento da métrica de autoridade, onde as listas ranqueadas τ_i e τ_j representam as listas ranqueadas de img_i e img_j e a lista τ_q é a lista da imagem de consulta img_q . As linhas azuis representam a co-ocorrência dos vizinhos recíprocos.

Figura 12 – Ilustração do grafo de vizinhança com ocorrência de vizinhos recíprocos.



Fonte: Adaptado de (PEDRONETTE; PENATTI; TORRES, 2014).

- **Reciprocal Neighborhood Density**

A medida de Densidade de Vizinhança Recíproca (*Reciprocal Neighborhood Density*) tem seu funcionamento como descrito a seguir: para cada imagem de consulta, ela examina as k imagens mais próximas (vizinhança), onde duas imagens são consideradas vizinhas se $img_i \in \mathcal{N}(q, k) \wedge img_q \in \mathcal{N}(i, k)$ e para determinar isso utiliza-se a função $f_r(j, l) \rightarrow \{0, 1\}$. Em seguida, verifica quantas dessas k imagens

também ranqueiam a imagem de consulta entre suas próprias k imagens mais próximas. A existência dessas relações recíprocas indica uma maior coesão e relevância dentro do conjunto de resultados. Um peso é atribuído a cada relação recíproca, definido pela função $w_r(q, j) = k + 1 - \tau_q(j)$, sendo maior para imagens que aparecem em posições mais altas nas listas ranqueadas.

A densidade final é calculada com base no número e nos pesos dessas relações recíprocas. Quanto maior a densidade, maior a probabilidade de que a lista de resultados seja efetiva. A fórmula considera a posição de cada imagem na lista ranqueada, dando mais peso às imagens de classificação mais alta (PEDRONETTE; TORRES, 2015). A função de $n_r(q, i)$ é expressa pela Equação 3.5:

$$n_r(q, i) = \frac{\sum_{i \in \mathcal{N}(q, k)} \sum_{i \in \mathcal{N}(i, k)} f_r(j, l) \times w_r(q, j) \times w_r(i, l)}{k^4} \quad (3.5)$$

Em resumo, a metodologia usa a informação recíproca entre as imagens de alta posição no ranking para inferir a qualidade da recuperação, sem depender de dados rotulados ou treinamento. A alta correlação observada com medidas de precisão estabelecidas indica sua eficácia na predição da efetividade de consultas em sistemas de recuperação de imagens.

3.4.2 Métricas de Correlação

Desempenham um papel fundamental ao identificar a força e a direção das relações entre os descritores e o resultado desejado. Ao calcular a correlação entre os descritores, as métricas de correlação revelam quais descritores possuem maior impacto ou influência nos resultados, auxiliando na escolha dos descritores mais relevantes e impactantes para a análise ou o modelo em questão.

Dentre as métricas de correlação pode-se destacar:

- **Coefficiente de Jaccard**

O coeficiente de Jaccard é uma métrica estatística que calcula a correlação entre duas listas ranqueadas e retorna um valor no intervalo de $[0,1]$, quanto mais próximo de 1 maior a correlação (LIN et al., 2015). É expresso pela Equação 3.6:

$$J(\tau_i, \tau_j) = \frac{|\tau_i \cap \tau_j|}{|\tau_i \cup \tau_j|} \quad (3.6)$$

- **Coefficiente de $Jaccard_k$**

O coeficiente de Jaccard estabelece tradicionalmente a relação entre o tamanho da intersecção sobre o tamanho da união em apenas uma profundidade k , assim ignorando os elementos abaixo das k primeiras posições.

Para amenizar isso Okada, Pedronette e Torres propuseram calcular o acumulado da pontuação de Jaccard, considerando diferentes pesos (OKADA; PEDRONETTE; TORRES, 2015). Essa métrica é expressa pela Equação 3.7:

$$J_k(\tau_i, \tau_j, k) = \frac{\sum_{d=1}^k J(\tau_i, \tau_j, k)}{k} \quad (3.7)$$

- **Rank Overlap Biased (RBO)**

A sobreposição baseada no ranqueamento (RBO) compara a sobreposição de dois elementos com profundidades crescentes. A profundidade de interesse pode ser variada ajustando um parâmetro denominado *persistência* (p) no intervalo de $[0,1]$ enquanto a probabilidade pelo usuário é dada por $1 - p$ para uma profundidade d . Quanto mais baixo o valor de p mais importância é dada aos elementos do topo, e quando $p = 0$ somente o primeiro elemento é considerado (SARICA; QUATTRONE; QUATTRONE, 2022). Essa abordagem é matematicamente expressa pela Equação 3.8:

$$RBO(\tau_i, \tau_j, k, p) = (1 - p) \sum_{d=1}^k p^{d-1} \times \frac{|\mathcal{N}(i, k) \cap \mathcal{N}(j, k)|}{d} \quad (3.8)$$

- **Correlação de Spearman**

A correlação de Spearman, denotada com a letra ρ , é uma métrica não paramétrica que avalia a relação entre pares de variáveis. Seu resultado se encontra no intervalo $[-1, 1]$, onde -1 indica uma correlação negativa perfeita, 1 indica uma correlação positiva perfeita e 0 indica ausência de correlação linear entre as variáveis (WISSLER, 1905; MYERS; SIROIS, 2004), formalmente expressa pela Equação 3.9:

$$S_\rho(\tau_i, \tau_j, k) = \frac{\sum_{x,y \in \mathcal{N}(i,k) \cup \mathcal{N}(j,k)} |\tau_i(j) - \tau_j(i)|}{2 \times k \times n} \quad (3.9)$$

Onde n é um domínio de tamanho fixo.

- **Kendal $_\tau$**

O τ de Kendal é uma apuração tradicional de distância entre permutações nas k primeiras posições via uma ordenação *Bubble sort* (FAGIN; KUMAR; SIVAKUMAR, 2003a; OKADA; PEDRONETTE; TORRES, 2015), expresso pela Equação 3.10:

$$K_\tau(\tau_i, \tau_j, k) = 1 - \frac{\sum_{x,y \in \mathcal{N}(i,k) \cup \mathcal{N}(j,k)} \bar{K}_{x,y}(\tau_i, \tau_j)}{k \times (k - 1)} \quad (3.10)$$

Nessa equação $\bar{K}_{x,y}(\tau_i, \tau_j)$ é a função que determina se os objetos img_x e img_y estão na mesma ordem em comparação as listas ranqueadas τ_i e τ_j , essa função é expressa pela Equação 3.11:

$$\overline{K}_{x,y}(\tau_i, \tau_j) = \begin{cases} 0 & \text{se } (\tau_i(x) \leq \tau_i(y) \wedge \tau_j(x) \leq \tau_j(y)), \\ 0 & \text{se } (\tau_i(x) \geq \tau_i(y) \wedge \tau_j(x) \geq \tau_j(y)), \\ 1 & \text{caso contrário} \end{cases} \quad (3.11)$$

Para a realização da função $\overline{K}$ a posição máxima k adotando $k + 1$ para posições acima de k .

4 Abordagens Recentes

Este capítulo visa apresentar as abordagens de recuperação da informação com foco em recuperação de imagens baseadas em conteúdo baseado em agregação tardia de listas ranqueadas, seguindo a estrutura de seções, sendo apresentadas abordagens baseadas em grafos na Seção 4.1, abordagens hierárquicas ou multinível na Seção 4.2, abordagens baseadas em redes neurais na Seção 4.3, abordagens aplicadas em outras áreas de recuperação da informação, mas que possuem potencial de aplicação em recuperação de imagem baseada em conteúdo na Seção 4.4, abordagens de agregação baseadas em subespaço, *framework* genético, modelos semi-supervisionados e fusão adaptativa na Seção 4.5 e bibliotecas de agregação na Seção 4.6.

4.1 Abordagens baseadas em Grafo

Os modelos de agregação baseados em grafos tem sido recorrentes na literatura, um modelo de agregação baseado em grafos permite a combinação de informações em um conjunto de dados que pode ser representando por meio de um grafo, onde os nós representam elementos e as arestas simbolizam as conexões entre os elementos. O conceito base de um modelo de agregação baseado em grafos é de que cada nó do grafo agregar informações de seus vizinhos para aprender ou aprimorar a representação de si próprio. Este procedimento ocorre por meio de diversas camadas de agregação, nas quais cada camada expande a "perspectiva" do nó sobre a rede, possibilitando a identificação de relações diretas e indiretas entre os nós (DOURADO; PEDRONETTE; TORRES, 2019; ŠKRLJ; REPAR; POLLAK, 2019).

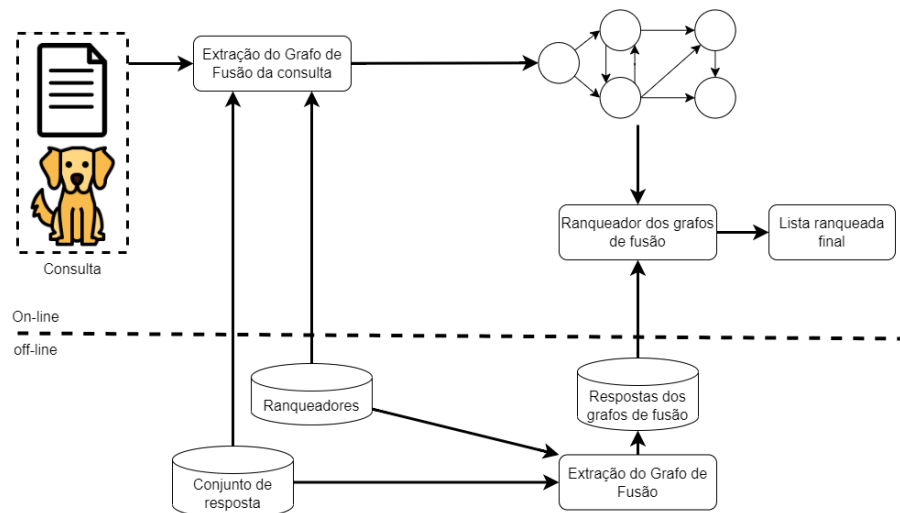
O modelo de agregação em grafos é usado para combinar resultados de modelos de ranqueamento isolados em tarefas de recuperação de informação. A metodologia é não supervisionada, independente da formulação dos ranqueamentos isoladas, podendo combinar modelos arbitrários baseados em critérios de ranquemaneto diferentes (textuais, imagens ou híbridos). Se baseia em duas etapas: Extração de grafos de fusão - esta etapa agrega listas ranqueadas para uma consulta, criando um grafo de fusão por consulta. Inclui a criação de listas ranqueadas usando diferentes ranqueadores, normalização das listas ranqueadas e fusão das listas ranqueadas; Ranqueador de grafos de fusão - tanto a consulta quanto cada amostra da coleção são representadas por grafos (grafo de fusão da consulta e grafo de fusão da amostra). A busca é modelada como o ranqueamento dos grafos de fusão das amostras da coleção em relação a um grafo de fusão da consulta. A similaridade entre os grafos é calculada utilizando o subgrafo comum mínimo.

O problema de recuperação ad-hoc (refere-se ao processo de buscar e recuperar

informações em resposta a uma consulta específica) é reformulado como recuperação de documentos baseada em grafos de fusão, uma nova representação unificada que mescla múltiplos rankings e expressa automaticamente as inter-relações dos resultados de recuperação. A hipótese é que o sistema de recuperação se beneficia do aprendizado da estrutura de variedade dos conjuntos de dados, levando a resultados mais eficazes.

Ao contrário de abordagens existentes, a formulação de agregação baseada em grafos proposta encapsula informações contextuais codificadas de múltiplos ranqueamentos, diretamente utilizáveis para o ranqueamento, sem cálculos adicionais ou etapas de pós-processamento nos grafos. Uma nova pontuação de recuperação de similaridade é formulada usando um cálculo eficiente de subgrafos comuns mínimos. Finalmente, uma vantagem sobre abordagens existentes é a ausência de hiperparâmetros (DOURADO; PEDRONETTE; TORRES, 2019).

Figura 13 – Esquemática de um sistema de agregação baseado em grafos.



Fonte: Adaptado de (DOURADO; PEDRONETTE; TORRES, 2019).

Uma abordagem que apresenta resultados relevantes é a fusão seletiva de rankings baseada em grafos, na qual as características são representadas como nós e as arestas representam as relações de complementaridade. Nesse modelo, o grafo é atualizado iterativamente, considerando diferentes limiares de medidas para estimar a eficácia. O objetivo é selecionar as melhores características, ou seja, aquelas que possuem maior eficácia (VALEM; PEDRONETTE, 2020a).

As arestas são ajustadas considerando os limiares das medidas de correlação de ranqueamento. As características selecionadas são determinadas pelos componentes conectados do grafo, os quais são combinados por meio de uma estratégia de *late fusion*. Os testes realizados com essa abordagem demonstraram ser promissores para uma seleção eficaz de combinações.

Existem modelos que se baseiam em matrizes, onde são considerados os M *rankings* das N alternativas dadas pelos M *rankings*, no caso da recuperação de imagens seriam considerados os descritores, assim a matriz de ranques é denotada por $R = ((r_{ij})_{M \times N})$, onde R_{ij} representa o *ranking* de alternativas a_j para cada descritor b_i , caso não haja alternativa a_j para o eleitor b_i então o valor de $r_{ij} = 0$. A matriz de transição dos *rankings* b_i é denotada por $P^i = (P_{st}^i)_{N \times N}$, onde $P_{st}^i = 1$ quando a alternativa a_s supera a alternativa a_t , e $P_{st}^i = 0$ caso contrário. Essa abordagem têm como base a matriz de competição, definida como $A = (a_{st})_{N \times N}$, onde $a_{st} = \sum_i^M P_{st}^i$, sendo adaptado para um grafo de competição definido por $G = (V, E)$, onde V é o conjunto de nós representando as alternativas, e $E \subseteq V \times V$ é o conjunto de arestas representando as gravações de perda-ganho entre as alternativas (XIAO et al., 2021).

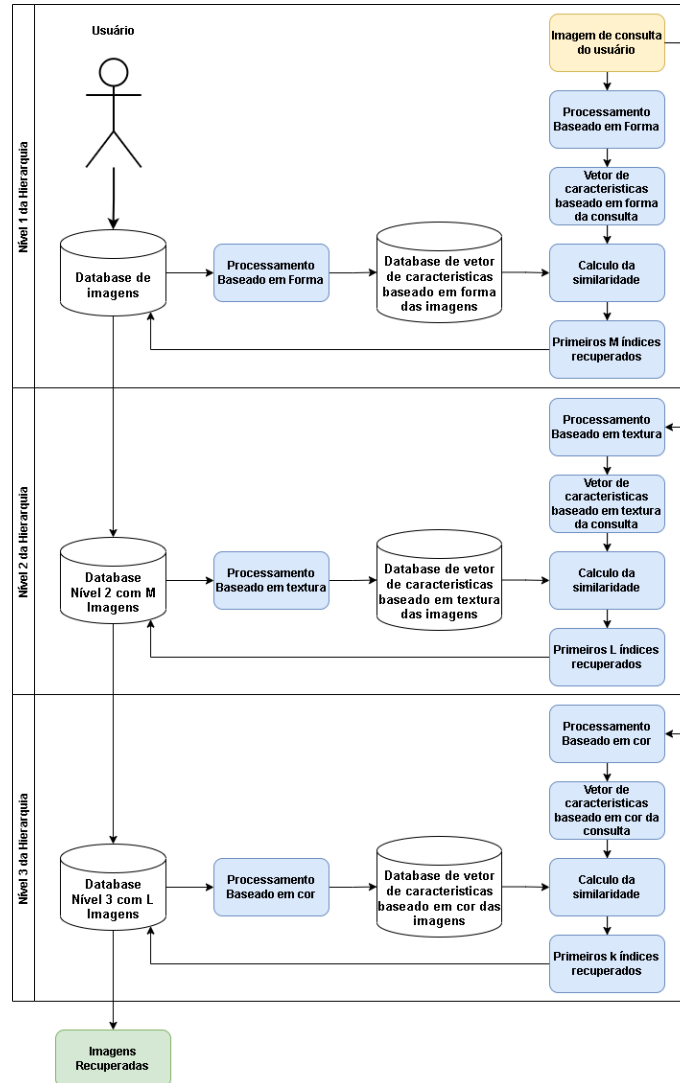
A abordagem de fusão de vetores proposta por Dourado e Torres (2019) foi desenvolvida para inovar funções de agregação não-supervisionada de listas ranqueadas, intrinsecamente multimodais, e presando por velocidade de recuperação e eficácia máxima.

Essa abordagem se baseia no modelo de grafos, entretanto utiliza um indexador de vetores de fusão, e a partir dos índices dos vetores de fusão é possível recuperar itens de resposta para uma consulta, desde que tanto os itens de resposta como os itens da consulta sejam previamente representados como vetores de fusão. O passo seguinte é o processo de ranquear os vetores de fusão, através da similaridade entre os vetores de fusão dos itens de resposta e o vetor de fusão do item de consulta.

Apesar de ser possível ao ranqueamento dos vetores de fusão diretamente, o processo de indexação é importante para promover um tempo de processamento de consulta sub-linear.

4.2 Abordagens de Agregação Hierárquica de Características e Abordagens de Agregação multinível

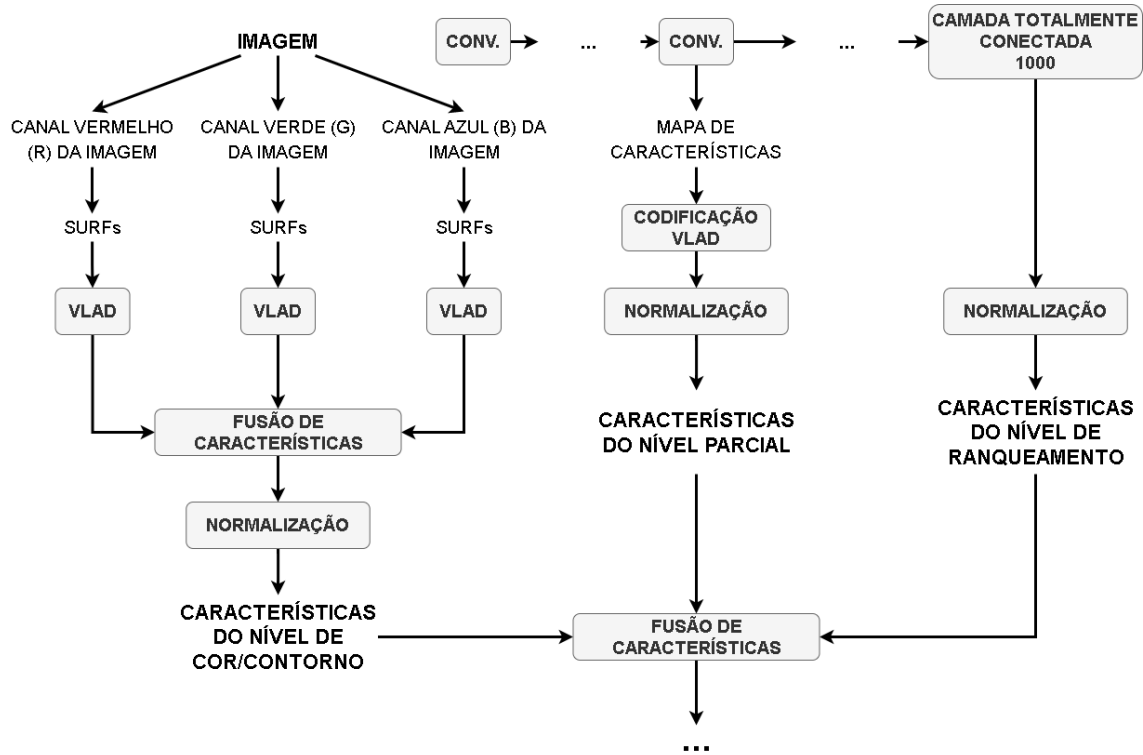
Os modelos baseados em CBIR tradicionais utilizam de características primitivas (cor, textura, forma/contorno, etc.) combinadas. Entretanto, esse tipo de fusão nem sempre é o mais eficaz, pois uma característica pode ofuscar a outra. Para minimizar esse problema pode-se utilizar um *framework* hierárquico, ou multinível, onde cada grupo de característica é considerado em uma única hierarquia, tendo como principal preocupação a seleção adequada da ordem de recuperação das características visuais. Para aplicação dessa metodologia o primeiro passo é dividir as hierarquias. Em Kumar et al. (2023) os autores empregam os níveis de forma, textura e cor, executando os descritores respectivos cascadeando o processo em fases, como visto na Figura 14.

Figura 14 – Diagrama de agregação multinível baseada em hierarquia.

Fonte: Adaptado de (KUMAR et al., 2023)

A abordagem multinível proposta em Wu e Yu (2019) se baseia em três níveis: nível de cor/contorno, nível parcial e nível cenário, como pode ser visto na Figura 15, cada nível é responsável por ações específicas, como o nível de cor/contorno que extrai as características de contorno e de cor e concatenar em um único descritor *vector of locally aggregated descriptors* (VLAD).

Figura 15 – Arquitetura do modelo de fusão multinível.



Fonte: Adaptado de (WU; YU, 2019)

O descritor local mais popular para descrever contornos e texturas é o SIFT, entretanto as características são extraídas de imagens em tons de cinza perdendo as informações de cor. Assim o modelo adotado pelos autores trata-se do SURF para cada canal de cor fundindo os três vetores em um único vetor de características, o nível parcial trata das camadas intermediárias da convolução, que tendem a capturar informação de partes de objetos e o nível cenário trata das camadas totalmente conectadas que tendem a capturar mais informação global e pode ser utilizada para representar o *layout* espacial da cena global. A partir dos vetores gerados por cada nível, uma nova fusão é efetuada agregando as características dos três níveis em um único vetor.

Os modelos de recuperação de imagens de fina granularidade, do inglês *Fine-grained image retrieval* (FGIR) têm recebido muita atenção, porém problemas de grandes e pequenas diferenças interclasse ainda se encontram em aberto. De modo a amenizar tais problemas, o modelo *coarse-to-fine multiple-level feature fusion network* (MFFN) foi proposto em Wang et al. (2022a).

Primeiramente o modelo proposto adota características ao nível de objeto para uma recuperação grosseira, ou seja, características de alto nível para reduzir o escopo de recuperação. Os autores utilizam o *Selective Convolutional Descriptor Aggregation* (SCDA) para essa etapa. Caso essa fase seja omitida, será necessário calcular características refinadas

para todas as imagens da base de dados, levando a uma alta complexidade computacional, o que não é propício para uma recuperação eficiente.

A fase de recuperação refinada são utilizadas características de trechos (*patches*) capazes de refletir as diferenças sutis entre os objetos, melhorando a precisão da recuperação. Para que os trechos selecionados não percam sua representatividade, uma vez que podem conter muito ruído enquanto os trechos realmente distinguíveis se encontram distribuídos na área do objeto, aplica-se uma restrição que incentiva a sobreposição dos trechos selecionados sobre o objeto, o máximo, assim, utiliza-se a sobreposição de *Intersection over Union* (IoU) a qual é amplamente utilizada no cálculo de sobreposição entre dois retângulos. Também é aplicada uma restrição aos trechos para haver o menor número de sobreposições possíveis, utilizando a sobreposição de IoU para isso.

A fase de fusão dos trechos de características ocorre geralmente através da fusão média de todos, porém o modelo de fusão média trata todas as características igualmente, assim caso a proporção dos trechos seja ampliado para incluir mais informação de fundo, a habilidade de caracterização do trecho de características final é reduzido. Para isso pode-se adotar o método agrupamento máximo ponderado (*weighted max-pooling*), pois ao reduzir significativamente a informação de fundo (*background*) redundante em cada trecho de características é possível reter informações complementares sobre os trechos descrevendo diferentes áreas do objeto, gerando assim um trecho com maior capacidade de distinção. Na fase de fusão de características complementares utiliza-se uma rede *Deep Belief Network* (DBN) e um *Restricted Boltzmann Machine* (RBM) em conjunto para agregar características multinível, se várias características são inseridas na rede DBN, aplica-se um RBM conjunto para fundir a representação obtida pelas DBNs, e a distribuição conjunta entre as características pode ser aprendida automaticamente, obtendo assim informações complementares ricas.

Para aprimorar a acurácia da recuperação, utiliza-se a fase de expansão de consulta através da soma e média das cinco primeiras características das imagens retornadas pela fase de recuperação de granulação refinada para calcular um novo descritor de consulta, então uma nova rodada de granulação refinada é executada com o novo descritor de consulta.

O esquema de correspondência em vários níveis proposto em Geetha et al. (2020) apresenta uma recuperação de imagens com base em uma similaridade de características híbridas que integram características locais e globais, onde os pontos de interesse de cada imagem são obtidos pelos descritores LBP e SURF, além da extração das características globais de cor que utilizam um novo descritor de direcionamento de borda e características baseadas em cor. Ambas as características globais e locais incluídas no esquema de multi nível são utilizadas para representar a imagem.

O método *Unsupervised Selective Rank Fusion* (USRF) (VALEM; PEDRONETTE,

2020c) consiste em uma estrutura não supervisionada para seleção e fusão de recursos visuais em tarefas de recuperação de imagens, projetada para otimizar a agregação de resultados provenientes de diversos sistemas de recuperação de informação (IR) ou algoritmos de ranqueamento. Ao invés de empregar uma simples combinação linear ou métodos de fusão baseados em aprendizado supervisionado, a USRF adota uma abordagem seletiva e adaptativa que identifica os rankings mais relevantes e combina seus resultados de maneira ponderada. Este método implementa mecanismos sofisticados para selecionar os rankings mais promissores a partir de um conjunto de rankings fornecidos, baseando-se em critérios como diversidade, complementaridade e relevância dos resultados. A metodologia visa superar as dificuldades de selecionar a melhor combinação de recursos visuais em cenários sem dados rotulados, operando em quatro etapas principais:

- **Estimativa de Eficácia**

Para cada ranqueador individual, a USRF calcula uma medida de eficácia usando métricas baseadas em ranking que não exigem dados rotulados, como o Authority Score e a Reciprocal Density. Essas métricas analisam a densidade de referências de ranking nas posições superiores das listas retornadas por cada ranqueador, estimando sua performance, representada na Figura 16 pelo número (1).

- **Medidas de Correlação de Ranking**

A USRF calcula medidas de correlação entre pares de ranqueadores usando métricas como Jaccard, $Jaccard_k$, RBO, Spearman e $Kendall_\tau$, representada na Figura 16 pelo número (2). Essas métricas quantificam a complementaridade entre os resultados de recuperação de cada par, identificando se eles fornecem informações distintas e, portanto, se sua combinação pode ser benéfica.

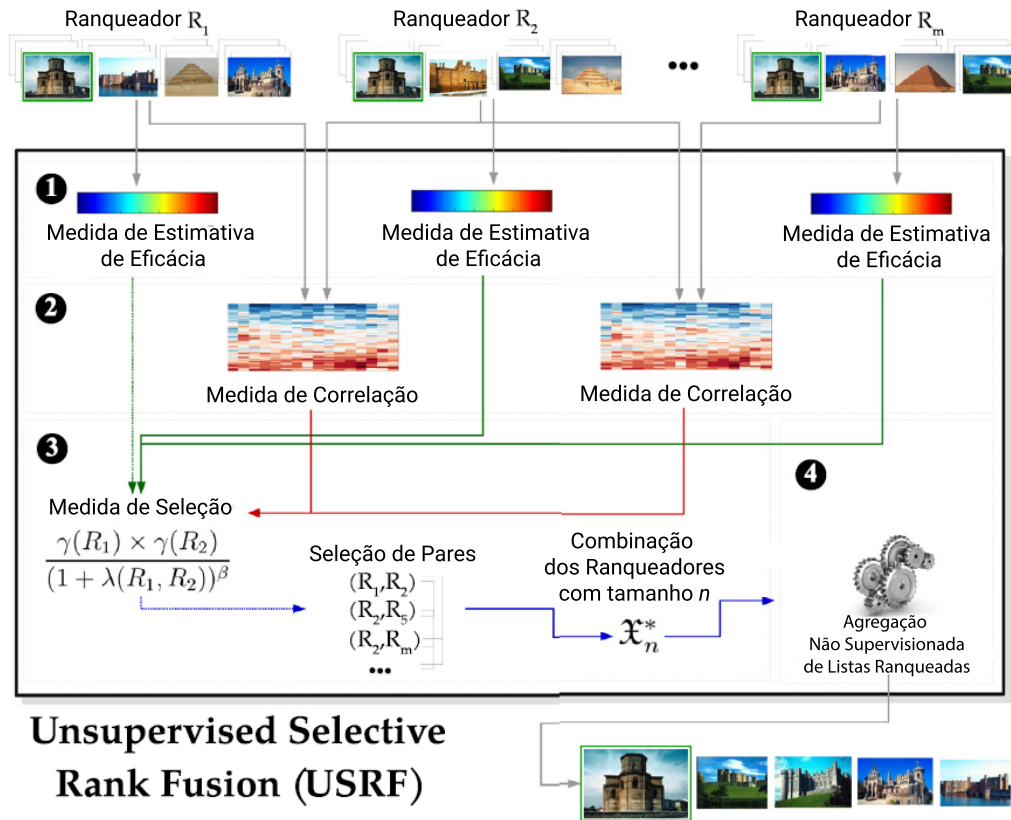
- **Estratégia de Seleção**

Em vez de avaliar todas as possíveis combinações de ranqueadores (um problema computacionalmente caro), a USRF utiliza uma estratégia baseada em pares pela combinação dos ranqueadores de tamanho n . A seleção é feita com base em uma medida que combina as estimativas de eficácia individual e as medidas de correlação entre pares, representada na Figura 16 pelo número (3). Isso permite a seleção de combinações de dois ou mais ranqueadores, reduzindo significativamente a complexidade computacional. Esse processo envolve a geração de combinações de tamanho n a partir de ranqueadores previamente avaliados por suas eficácias individuais e pela correlação entre eles. As combinações são formadas iterativamente, partindo de combinações menores ($n - 1$) e ampliando-se até o tamanho n , garantindo que a combinação final maximize a eficácia geral

- **Agregação de Ranking**

Os ranqueadores selecionados na etapa anterior são então combinados usando um método de agregação de ranking. Este passo gera a lista final de recuperação de imagens, que representa a saída do método USRF, representada na Figura 16 pelo número (4).

Figura 16 – Ilustração Método USRF.



Fonte: Adaptado de (VALEM; PEDRONETTE, 2020c)

4.3 Abordagens baseadas em Redes Neurais

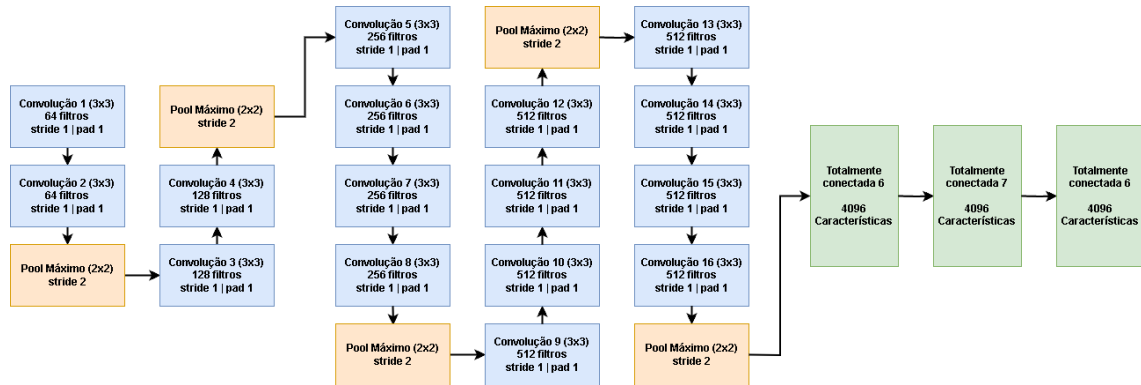
Modelos baseados em redes neurais podem executar a fusão precoce e tardia, dependendo do momento da captação das listas ranqueadas, podendo ser executada nas camadas de convolução ou ao final de todas as camadas.

O modelo utilizado em Pathak e Raju (2021) propõe uma melhor eficácia na recuperação de imagens através da utilização de um *framework* eficiente para recuperação de imagens por meio de fusão de características de alto nível da *GroupNormalized-Inception-Darknet-53* e características clássicas extraídas dos modelos de cores RGB e *Hue, Saturation and intensity* (HSI), e de contorno baseado em gradiente (HoG), ou seja, a fusão de características clássicas e características obtidas por meio de uma CNN. Para extração

mais detalhada das imagens, os autores aumentam uma camada de inicialização da CNN, incluindo os núcleos (*kernels*) 1×1 e 5×5 ao núcleo 3×3 existente. Visando reduzir a dependência do tamanho do lote, os autores utilizaram a camada de normalização em grupo (*Group Normalization* (GN)) ao invés de uma camada de organização em lote (*Batch Normalization* (BN)). Uma versão modificada da codificação de truncamento de blocos com difusão de pontos (*dot-diffused block truncation coding* (DDBTC)) para extração das características no espaço RGB, assim o processo de fusão é realizado a partir do modelo de extração proposto.

A abordagem utilizando a arquitetura VGG-19 consiste em 16 camadas de convolução e três camadas totalmente conectadas. Essa arquitetura utiliza filtros 3×3 para todas as camadas, visando aprender o máximo de características possíveis. Após as camadas de agrupamento (*pooling*), o número de filtros é dobrado para aumentar a profundidade, enquanto as dimensões espaciais são mantidas. As imagens coloridas, com proporção de 224×224 , passam por um pré-processamento em que é extraída a mediana dos valores RGB. Em seguida, as imagens seguem para as camadas de convolução com um passo (*stride*) e preenchimento (*pad*) de um *pixel*. As dimensões das imagens são reduzidas pela metade por meio de cinco camadas de agrupamento máximo (*pooling*) com tamanho dois e passo 2, sem preenchimento ocasionalmente entre as camadas, conforme pode ser observado na Figura 17 (BIBI et al., 2022).

Figura 17 – Arquitetura do modelo VGG-19.



Fonte: Adaptado de (BIBI et al., 2022)

Strides são parâmetros que determinam o deslocamento do filtro durante a convolução e afetam o tamanho de saída, além de poderem ser ajustadas para controlar a complexidade computacional e a resolução das características extraídas, enquanto *pads* referem-se ao preenchimento adicionado às imagens de entrada para controle do tamanho de saída após a convolução.

Para refinar o vetor de características descartando informações irrelevantes e redundantes aplica-se um algoritmo genético calculando a probabilidade de cada cromossomo

através dos K vizinhos mais próximos, e aplicando o método de roleta para seleção dos genes para mutação e *crossover*, onde os genes com melhor aptidão ocupam uma fatia maior da roleta proporcionando a seleção aleatória das características, porém respeitando a aptidão obtida por cada característica individualmente.

O modelo baseado em CNN proposto em Wang et al. (2022b) trata do uso de uma arquitetura VPR (*Visual Place Recognition*) inovadora baseada em *transformers* que combina detalhes locais de baixo nível, contexto espacial e informações semânticas de alto nível para uma melhor robustez em relação a mudanças de aparência e ponto de vista. Em síntese, trata-se de uma rede híbrida de extração de características *CNN-Transformer* que aprende o padrão visual considerando os detalhes locais e as dependências globais. Uma técnica de fusão em múltiplas escalas também foi desenvolvida que gera descritores locais-globais em vários níveis a partir da estrutura hierárquica do híbrido *CNN-Transformer* e, em seguida, integra pistas semânticas em várias escalas por meio de uma estratégia de agregação semântica, resultando em um melhor desempenho se comparado aos métodos de escala única.

Esse modelo híbrido baseia-se na arquitetura de rede neural *Swin Transformer*, que se trata de uma rede proposta para tarefas de visão computacional, pois ao contrário das arquiteturas convolucionais tradicionais, o *Swin Transformer* se baseia no mecanismo de atenção das *Transformers*, que originalmente era utilizada em processamento de linguagem natural (LIU et al., 2021). Dessa maneira, o modelo híbrido proposto beneficia-se do mecanismo de atenção e janela deslizante, que se trata da abordagem de dividir as imagens sobrepostas, ou janelas, e processá-los separadamente, o *Swin Transformer* pode modelar eficientemente dependências de longo alcance entre características locais de imagens, que se refere às relações entre diferentes regiões das imagens, assim como agregar automaticamente pistas visuais referentes à tarefa. Tais características podem ser extremamente positivas em cenários com mudanças de ponto de vista, compensando a deficiência da CNN em concentração em informações globais.

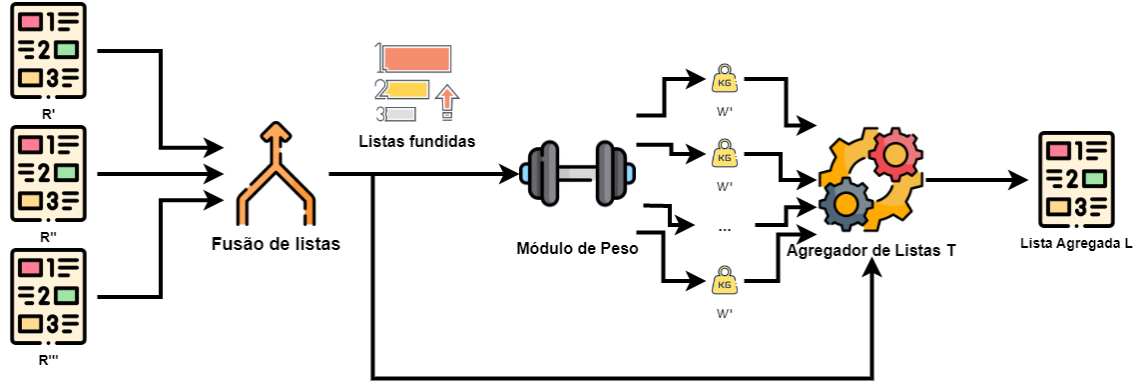
4.4 Abordagens de Outras Áreas de Agregação com Potencial para Aplicação em Recuperação de Imagens

Em Akritidis et al. (2022) é proposta a introdução de um modelo de algoritmo de agregação de listas ranqueadas não-supervisionado que automaticamente qualifique o grau de perícia de cada ranqueador através do uso de pesos calculados do eleitor, que é atualizado a cada iteração, executando a medida das distâncias entre cada lista e uma lista de agregação temporária. Diferente dos métodos baseados em *score*, essa abordagem considera somente o *ranking* individual dos elementos da lista de entrada.

Essa abordagem possui dois benefícios se comparada as demais metodologias

propostas até o momento, sendo que a utilização de listas de agregação temporárias evita problemas otimização e também é menos suscetível a problemas locais quando duas ou mais listas ranqueadas de baixa qualidade possuem alta semelhança. Para o funcionamento desta metodologia utiliza-se a função de distância denominada *Cosine Distance for Rank Aggregation* (CODRA), originado da bem estabelecida métrica de similaridade do cosseno. O fluxo desta abordagem pode ser visto na Figura 18.

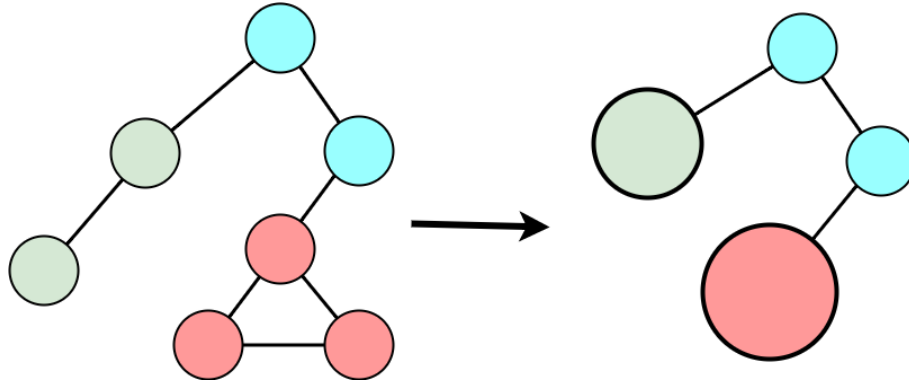
Figura 18 – Fluxo da abordagem baseada em pesos.



Fonte: Adaptado de (AKRITIDIS et al., 2022)

Para a Figura 18 adota-se V como o conjunto de n ranqueadores, cada ranqueador $v \in V$ aceita uma consulta q e gera uma lista ordenada $R^v = \{r_1^v, \dots, r_k^v\}$ onde $k = |R^v|$ elementos que suprem a consulta q . A notação r_k^v define a i -ésimo elemento na lista de entrada de v . O agregador utilizado é denotado por τ , W como o peso de cada ranqueador e $L = \tau(w_{v_1}, w_{v_2}, \dots, w_{v_n}, R^{v_1}, R^{v_2}, \dots, R^{v_n})$ é tido como a lista de pesos e suas listas ranqueadas para cada ranqueador $v \in V$, respectivamente.

O modelo de aprendizado não supervisionado e agregação por meta vértice proposto em Škrlj, Repar e Pollak (2019), mesmo tendo como proposta o ranqueamento de textos, sua estrutura de agregação por meta vértice pode ser promissora se aplicada em cenários de recuperação de imagens ao reduzir o número de vértices, assim como o de arestas. Isso envolve a fusão dos vértices candidatos em um meta-vértice que engloba o conteúdo de ambos os vértices candidatos, diminuindo a complexidade das conexões ao reduzir a quantidade de arestas necessárias para a ligação, como ilustrado na Figura 19:

Figura 19 – Fluxo de agregação de metavértices.

Fonte: Adaptado de (ŠKRLJ; REPAR; POLLAK, 2019)

Os passos de criação do meta vértice se baseiam nos passos de identificação dos meta-vértices candidatos, criação dos meta-vértices, para adaptação em listas ranqueadas pode-se dizer que cada lista contida no meta-vértice são reconectadas ao mesmo, as duas listas originais são removidas do grafo, e o procedimento é repetido para todos os candidatos. Com a devida adaptação esse modelo é promissor para otimização dos modelos de grafo convencionais.

Para a construção do meta vértice tem-se V como o conjunto de vértices, o meta vértice M é composto pelo conjunto de vértices de V , ou seja $M \subseteq V$, tomando M_i como a denotação do i -ésimo meta vértice. Na construção do meta-vértice M_i de modo que para cada $u \in M_i$, das u arestas iniciais são religadas a um novo M_i adicionado.

O modelo proposto por Qi, Yang e Tejedor (2020) para reconhecimento de idioma baseado em redes neurais (DNN) possui um grande potencial para utilização no cenário de recuperação de imagens. Essa abordagem utiliza da função de cálculo de divergência denominada de *Lovasz Bregman* (LB) para agregação das listas ranqueadas baseado em pontuação (*score*). Essa função foi derivada da divergência generalizada de Bregman, parametrizado por Lovasz adicionando uma função sub-modular, que pode ser definida pela propriedade de retorno decrescente.

A função LB pode ser aplicada em cenários de divergência entre permutações baseadas em pontuação e permutações baseadas em ordem. Para medir a qualidade e eficácia do algoritmo de ranqueamento utiliza-se o Ganho cumulativo descontado normalizado, do inglês *Normalized Discounted Cumulative Gain* (NDCG).

Outra abordagem que apresenta potencial para aplicação no cenário de recuperação de imagens é o modelo baseado em algoritmo genético proposto em Ahmad, Pal e Ganivada (2022), o qual se baseia no processo de seleção natural onde as soluções adequadas (cromossomos) são selecionados da população baseado em uma função de aptidão. Tais

cromossomos selecionados produzem descendentes que possuem uma melhor chance de sobrevivência devido às características herdadas dos cromossomos pais. O conceito do algoritmo genético se caracteriza como um conceito elitista, onde os melhores resultados são memorizados.

A representação de um cromossomo é baseado em duas características principais, sendo elas: (1) A posição única de um elemento indica sua posição na lista, portanto um elemento pode apenas aparecer em uma lista, (2) Todos os elementos devem estar presentes em uma lista. A partir desses dois fatos pode-se afirmar que o comprimento de um cromossomo é igual à composição registrada. Essa abordagem se trata de uma técnica de refinamento do resultado obtido pela agregação convencional utilizando algoritmos de agregação de listas ranqueadas ao utilizar a lista τ obtida da agregação para aplicar o algoritmo genético, aperfeiçoando o resultado obtido, por esse motivo tal abordagem pode ser aplicada facilmente ao cenário de recuperação de imagens.

4.5 Outras Abordagens

Um modelo que se mostrou altamente competitivo na área de fusão de características para recuperação de imagens baseadas em conteúdo foi o modelo que adota a fusão adaptativa via normalização de entropia cruzada para melhorar a complementariedade dos recursos heterogêneos. Esse mecanismo combina os recursos de forma não supervisionada e apresenta melhor resultado do que a recuperação por meio de um único recurso, assim a relação entre o peso de fusão e a complementariedade de características pode ser realizada adaptativamente (MA et al., 2023).

A abordagem multilinear de fusão de múltiplos índices utilizado em Zhang et al. (2019) propõe a utilização de uma matriz funcional para atualização dos índices originais, denominada de matriz funcional específica do índice, para propagação de similaridade. As matrizes funcionais são então otimizadas em um espaço tensor unificado para alcançar um refinamento, de modo a aproximar as imagens relevantes. O diferencial desta abordagem se encontra no uso de estruturas de subespaço de múltiplos índices nos novos índices como uma restrição difundida, garantindo assim pouco consumo adicional de memória na fase de consulta *on-line*.

Uma abordagem que apresenta um grande potencial se baseia em esboços como imagem de consulta, para captura de detalhes visuais de fina granularidade. Esse modelo se baseia na importância de lidar com características únicas dos esboços em oposição às imagens, a partir dos sucessos anteriores em modelos de recuperação de imagens baseada em esboços de baixa granularidade (*fine-grained sketch-based image retrieval* (FGSBIR)). O modelo proposto em Sain et al. (2020) visa o estudo de características negligenciadas até o momento sendo os níveis de detalhes, pois uma pessoa normalmente esboça em vários

graus de detalhamento para representar um objeto. Através do uso de uma rede capaz de cultivar hierarquias específicas de esboços e explorá-las a fim de efetuar correspondências entre esboço e imagem em níveis hierárquicos correspondentes.

Em específico, as características de um esboço e uma imagem são enriquecidas através do uso da co-atenção cruzada-modal juntamente com a fusão hierárquica de nós em cada nível para formação de um espaço de incorporação melhor para a recuperação.

Visando suprir os cenários de fusão multi-visualização não supervisionada, onde um subespaço latente compartilhado é recuperado da característica profunda original e da representação manual rasa, o modelo de fusão baseado em multi-visualização de subespaços foi utilizado em Li et al. (2021a). O processo de fusão das características se baseia na estratégia de subespaços, onde assumi-se que existe um subespaço de baixa dimensão subjacente aos recursos de múltiplas visualizações, criando um vetor de características compacto para representação da imagem. Esse modelo de aprendizado por subespaço pode ser combinado ao aprendizado discriminativo aumentando o poder discriminativo do subespaço embutido. Com a ajuda de modelagem de *rankings* de baixo nível (*low-rank*) e representação difundida, os subespaços recuperados utilizam da robustez desejada para o ruído embutido na multi-visualização enquanto mantém a máxima discriminação de características.

Existem modelos que propuseram a abordagem semi-supervisionada, mesmo estes modelos não sendo o foco principal deste trabalho, ainda, sim, é um modelo interessante de se exemplificar o resultado do uso de metodologias não supervisionadas juntamente a outras metodologias. Como a abordagem proposta em Pedronette et al. (2019), no qual aprimora o resultado obtido por metodologias não-supervisionadas por meio de realimentação de relevância semi-supervisionada. Neste modelo o aprendizado não-supervisionado ocorre por uma função que itera pelos elementos da coleção sem a necessidade das imagens estarem etiquetadas (*labels*).

O passo semi-supervisionado de realimentação de relevância executa quatro passos principais, sendo eles: (1) Apresenta-se uma pequena coleção de imagens para o usuário, (2) Coleta-se o *feedback* do usuário, e assim defini-se as imagens relevantes e não relevantes, (3) Utiliza-se o *feedback* obtido para aprender a necessidade do usuário e (4) Com os dados aprendidos recupera-se a melhor coleção de imagens para o usuário para ser utilizada na próxima iteração.

4.6 Bibliotecas de Agregação Não Supervisionada de Listas Ranqueadas

Vale ressaltar após a apresentação das abordagens acima a biblioteca de código aberto *Flexible high-performance library for Rank Aggregation* (FLAGR) apresentada em Akritidis, Alamaniotis e Bozanis (2023) que proporciona a aplicação de agregação de listas ranqueadas de forma simplificada e gera como resultado a lista de agregação e um arquivo de avaliação, ambos na extensão CSV. Possui em sua biblioteca métodos como, por exemplo, contagem de Borda (GARCÍA-LAPRESTA; MARTÍNEZ-PANERO; MENESES, 2009), CombSUM e CombMNZ (RENDA; STRACCIA, 2003) e técnicas baseadas na cadeia de Markov (SCULLEY, 2007).

A entrada é constituída por dois objetos, sendo eles os parâmetros de entrada que armazenam parâmetros vitais para execução como o algoritmo de agregação de listas ranqueadas, seus hiper-parâmetros, localização dos arquivos de entrada e saída, e os dados de entrada (listas ranqueadas) que contemplam a lista de preferências de entrada organizadas em um vetor de objetos de consulta. Tais objetos de consulta, por sua vez, são compostos por dois componentes primários, o agregador, que trata de um objeto universal que armazena em um vetor com a entrada das listas de preferências, aplicada o algoritmo selecionado e gera uma lista ranqueada de saída. O segundo componente é o avaliador, um objeto que calcula a qualidade da lista ranqueada gerada através do método de avaliação definido pelo usuário.

Outra biblioteca que pode ser citada é a *Unsupervised Distance Learning Framework* (UDLF) como características dessa biblioteca se destacam: (1) sete métodos já implementados, (2) facilidade de implementação de novos métodos, (3) leitura/exportação de arquivos em diferentes formatos de entrada/saída, (4) fácil utilização do arquivo de configuração, (5) compatível com diferentes sistemas operacionais e (6) código aberto (*open-source*) (VALEM; PEDRONETTE, 2017).

Dentre os métodos implementados à biblioteca UDLF, pode-se destacar:

- **Ranked Lists Similarities (RL-SIM)** Este método foi proposto para realizar agregação ou re-ranqueamento, levando em consideração a similaridade entre listas ranqueadas para a caracterização de informações contextuais em sistemas baseados em CBIR. Essa metodologia leva em conta exclusivamente a similaridade entre as listas ranqueadas, independentemente da similaridade entre as imagens (PEDRONETTE; TORRES, 2013).
- ***Cartesian Product of Ranking References (CPRR)***

O objetivo principal deste método é maximizar as informações de ranqueamento disponíveis. Usando operações de produto cartesiano sobre conjuntos de vizinhança,

essas operações criam novas relações de pares, utilizadas para encontrar informações subjacentes de similaridade. Podemos dividi-lo em: normalização de *rankings*: o processo de avaliação de referências de ranqueamento recíproco para melhorar a simetria das vizinhanças e, portanto, aumentar a eficácia das listas ranqueadas. Produto cartesiano de referências do ranqueamento: onde os conjuntos de vizinhança dos primeiros k vizinhos e os k vizinhos reversos são levados em consideração para calcular o produto cartesiano. Esses resultados são utilizados para determinar a medida de similaridade iterativa (VALEM; PEDRONETTE; ALMEIDA, 2018).

- ***Log-based Hypergraph of Ranking References (LHRR)***

Através da aplicação de uma função logarítmica, são atribuídos pesos às referências de classificação que formam a base para a representação dos hipergrafos e suas respectivas hiperarestas. Uma das vantagens do uso de hipergrafos é a capacidade das hiperarestas de se conectarem com qualquer número de vértices, permitindo a representação de relações de similaridade entre conjuntos de objetos, em oposição às abordagens baseadas em grafos que se limitam a representar relações entre pares de objetos. Outra vantagem a ser destacada é a independência em relação às medidas de distância, bem como a viabilidade de soluções algorítmicas eficientes (PEDRONETTE et al., 2019).

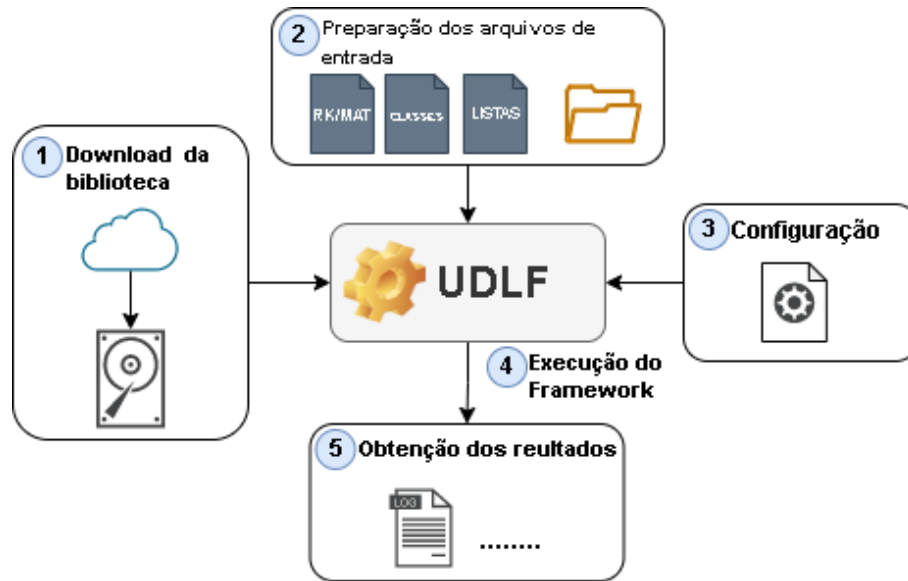
- ***Rank Flow Embedding (RFE)***

Essa abordagem é baseada em uma formulação de ranqueamento usada para criar um hipergrafo que pode representar informações de alta ordem de similaridade codificadas em listas ranqueadas. O foco desta metodologia é encontrar a melhor medida de similaridade para recuperação não supervisionada e uma incorporação mais eficaz para cada imagem (VALEM; PEDRONETTE; LATECKI, 2023b).

- ***Rank Diffusion with Assured Convergence (RDPAC)***

Esse método proporciona uma aproximação eficaz de um processo de difusão com base em dados de ranqueamento. Essa abordagem combina as vantagens dos métodos de difusão convencionais enquanto mantém a eficiência ao usar apenas as primeiras posições da lista ranqueada, enquanto assegura sua convergência. Para uma redução de custos computacionais, são empregadas matrizes de afinidades esparsas, calculando apenas um pequeno subconjunto de afinidades (PEDRONETTE; VALEM; LATECKI, 2021).

Figura 20 – Passos iniciais do UDLF.



Fonte: Adaptado de (VALEM; PEDRONETTE, 2017)

Na Figura 20 pode ser observado que o processo se divide em cinco passos iniciais, sendo eles a obtenção dos *softwares* da biblioteca UDLF através do site oficial e também pelo repositório oficial do GitHub, atualmente a biblioteca UDLF possui uma versão compatível com Python denominada PyUDLF.

O segundo passo é a preparação dos arquivos de entrada que necessita das listas/matrizes ranqueadas especificando o ranking de cada elemento da coleção ou a similaridade entre os elementos do *dataset*, e as listas especificando os nomes de cada elemento na coleção. Além dos arquivos obrigatórios pode-se adicionar os arquivos de classes quando o usuário está em busca das medidas de eficácia, e o caminho do *dataset* importante para coleções de imagens e caso o usuário queira visualizar a lista ranqueada em um arquivo HTML.

Terceiro passo é a configuração dos parâmetros como tarefa desejada, método que será utilizado, informações do *dataset*, arquivos de entrada e saída, configurações de avaliação, entre outros detalhes, configurado através da sintaxe de chave-valor *PARAMETER=VALUE*. Já os passos quatro e cinco tratam da aplicação da biblioteca e a obtenção dos resultados, respectivamente.

5 Metodologia de Agregação Não Supervisionada em Dupla Camada (UDLA)

Numerosas técnicas foram propostas para a extração de características visuais ao longo das últimas décadas, seja utilizando abordagens tradicionais ou baseadas em aprendizado profundo. Haja vista a ampla gama de métodos disponíveis, tarefas de fusão ou combinação visando aumentar a eficácia dos resultados consiste uma estratégia relevante na área. Contudo, a seleção e combinação de características ainda enfrenta vários desafios, especialmente em cenários não supervisionados, em que não há nenhum tipo de rótulo disponível. Dentre os desafios de pesquisa, pode-se destacar a dificuldade em selecionar as características a serem combinadas e os métodos de fusão em cenários não supervisionados. Nestes cenários, é necessário lidar com quantias numerosas de combinações possíveis sem o suporte de dados rotulados para direcionamento das escolhas.

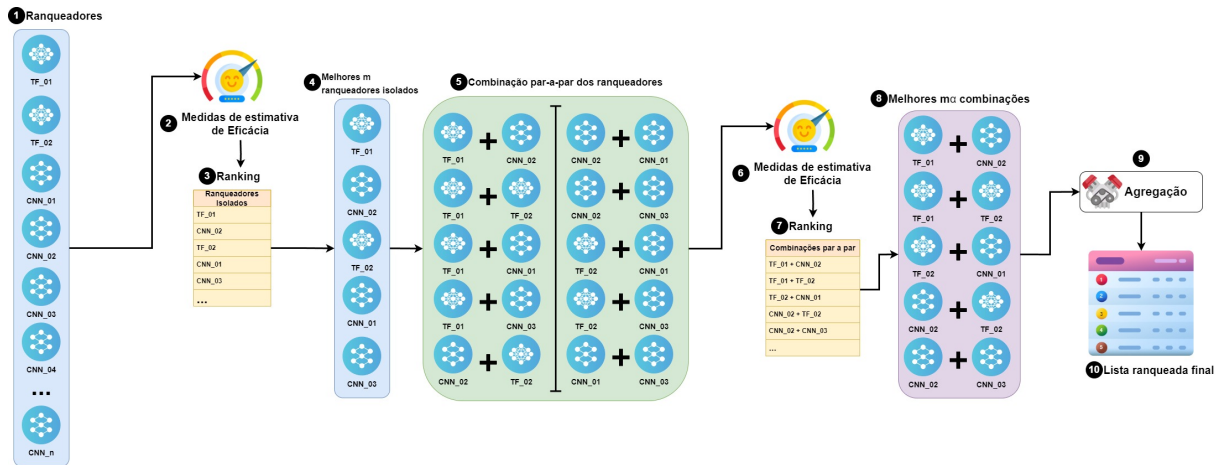
Esse trabalho apresenta uma metodologia para seleção e agregação de características visuais que pretende abordar os desafios mencionados, denominada de *Unsupervised Dual-layer Aggregation* (UDLA) ou agregação não supervisionada em camada dupla. A metodologia proposta baseia-se em estratégias de fusão tardia (*late fusion*) e agregação de listas ranqueadas. Outro aspecto fundamental consiste na caracterização multinível da proposta, que se concentra na combinação de duas camadas de agregação de listas ranqueadas, visando abordagens distintas em cada camada, possibilitando a complementariedade entre métodos distintos. A motivação para a escolha do modelo em dupla camada se baseou na possibilidade do aproveitamento da complementariedade de modelos de agregação baseados em precisão e revocação, uma vez que a precisão aplicada na primeira camada garante precisão maior dos resultados durante a criação dos pares combinados e ao aplicar revocação na segunda camada permite a complementariedade dos resultados dos pares combinados para agregação uma lista ranqueada única. Cada um dos desafios discutidos são abordados na proposta conforme discussão a seguir:

- **Seleção das Características:** os ranqueadores são selecionados por meio de medidas de estimativa de eficácia não supervisionadas baseadas em ranqueamento. Tais medidas atribuem um valor numérico que permite estimar os melhores ranqueadores.
- **Número de Combinações:** uma vez selecionados um sub-conjunto de m ranqueadores, eles são organizados em pares visando manter reduzido o número de combinações analisadas. Após a combinação pareada, a abordagem usa medidas de estimativa de eficácia também para selecionar os pares, mantendo a abordagem robusta a possíveis outliers isolados.

- **Métodos de Agregação:** A abordagem proposta utiliza métodos contextuais de agregação de listas em ambas as camadas, mas têm um enfoque em precisão na primeira camada e em revocação na segunda camada.

Este capítulo apresenta e discute em detalhes a abordagem proposta, cuja estrutura fundamental é baseada na agregação multinível de ranqueadores. A proposta é ilustrada na Figura 21. Nas próximas seções, as etapas serão apresentadas de forma detalhada, sendo as etapas 1, 2 e 3 abordadas na Seção 5.1, as etapas 4 e 5 na Seção 5.2, as etapas 6, 7 e 8 na Seção 5.3 e as etapas 9 e 10 abordadas na Seção 5.4. Com base nas medidas de estimativa de eficácia, o método realiza a seleção não supervisionada dos m ranqueadores mais bem posicionados. Cada par de elementos entre os top- m , ranqueadores selecionados pela etapa anterior, é combinado com seus resultados focados em precisão através da aplicação da metodologia de agregação RDPAC (PEDRONETTE; VALEM; LATECKI, 2021), gerando novas listas ranqueadas. Estas listas são usadas para um novo processo de estimativa de eficácia e para a seleção das novas m_α , dada pela seleção de α em um intervalo de $]0, 1]$, multiplicado pelo total de elementos gerados na etapa anterior. Por fim, na última etapa, uma lista ranqueada final focada em revocação através do uso da metodologia de agregação CPRR (VALEM; PEDRONETTE; ALMEIDA, 2018) é criada através da agregação das listas ranqueadas geradas pelo processo da primeira camada.

Figura 21 – Organização geral da metodologia de agregação de ranqueadores multinível.



Fonte: Autor

5.1 Seleção de Ranqueadores

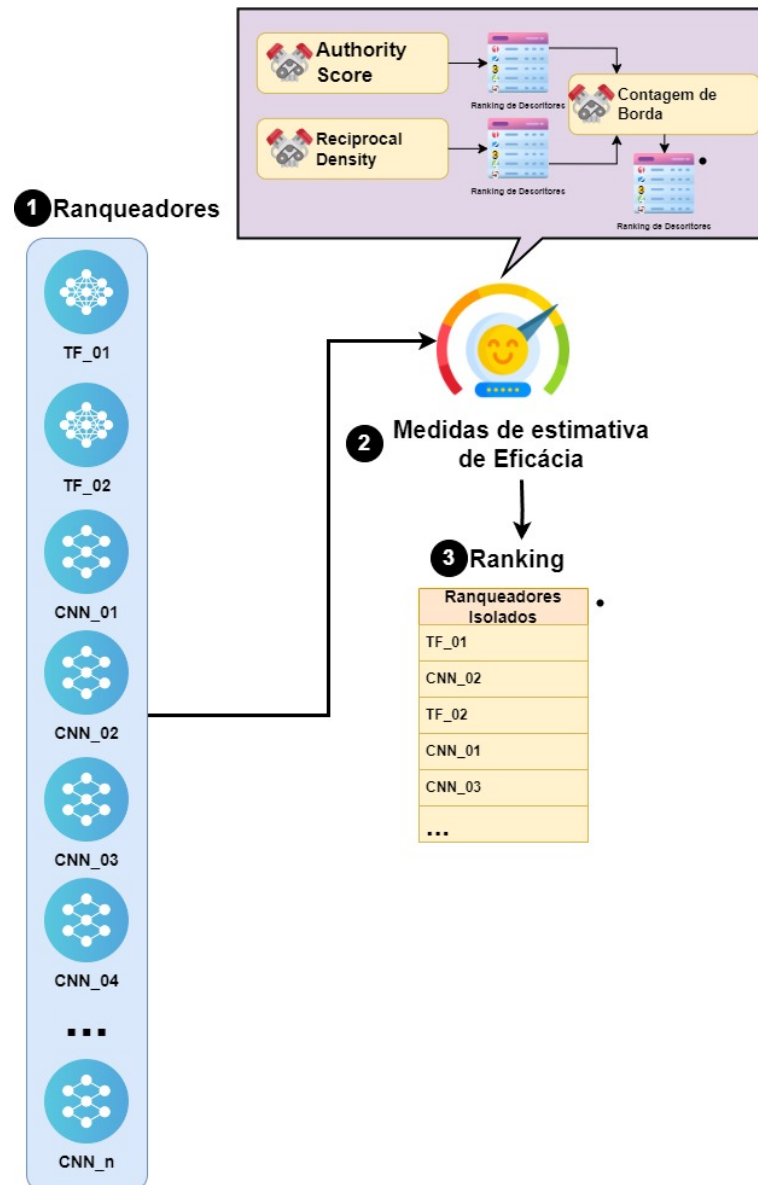
Um dos maiores obstáculos enfrentados no processo de combinação não supervisionada consiste na estratégia de seleção dos ranqueadores. Na ausência de dados rotulados, selecionar torna-se um relevante desafio. A abordagem proposta utiliza medidas de esti-

mativa de eficácia baseada em ranqueamento para esse propósito. Trata-se de técnicas que analisam a densidade da referência recíproca de ranqueamento, visando estimar quais ranqueadores apresentarão melhor resultados. Dentre as medidas disponíveis, destaca-se o *Authority Score*, cuja função é analisar a co-ocorrência dos elementos ranqueados nas k primeiras posições, onde k é tido como a profundidade da análise (PEDRONETTE; PENATTI; TORRES, 2014).

Destaca-se também o *Reciprocal Score*, que assim como o *Authority*, avalia a co-ocorrência, porém considera a posição em que se encontra em cada lista ranqueada, definindo um valor mais alto para elementos com co-ocorrência em posições mais altas nos *rankings* (PEDRONETTE; TORRES, 2015). Com o objetivo de explorar as informações de ambas as medidas, é executada uma agregação das listas obtidas pelos métodos *Authority Score* e *Reciprocal Density* utilizando o Método *Borda Count*, que se baseia na soma das posições dos rankings dados por cada medida. A motivação para o uso dessa abordagem foi a observação de que, embora apresentem boas estimativas, as metodologias *Authority* e *Reciprocal* produzem uma pequena variação entre os resultados obtidos. Dessa forma, com o objetivo de gerar uma lista mais precisa, foi aplicada a agregação pela contagem de Borda, que combinou os resultados das duas métricas de eficácia. Isso possibilitou a obtenção de uma lista ordenada que reflete o melhor das duas estimativas.

A Figura 22 ilustra o fluxo do processo de definição e seleção dos ranqueadores de forma não supervisionada, com cada etapa numerada.

Figura 22 – Ilustração do processo de seleção de ranqueadores.



Fonte: Autor

- **Etapa 1. Definição de ranqueadores isolados:**

Essencialmente essa etapa conta com a definição dos ranqueadores de entrada para o processo de agregação proposto;

- **Etapa 2. Estimativa de eficácia:**

Nessa etapa são aplicadas as medidas de estimativa de eficácia *Authority Score* e o *Reciprocal Score*, após obter o resultado dos ranqueadores ordenados de forma decrescente, aplica-se a agregação dos resultados de ambas as métricas através da contagem de Borda (GARCÍA-LAPRESTA; MARTÍNEZ-PANERO; MENESES, 2009), gerando uma nova lista ranqueada de ranqueadores.

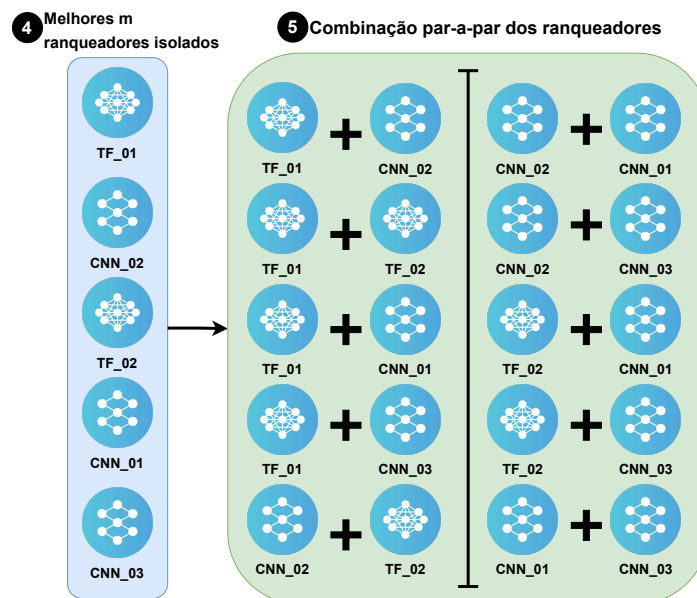
- **Etapa 3. Lista ranqueada de ranqueadores isolados:**

Após a aplicação da estimativa de eficácia, o objeto de saída desse processamento é a lista de ranqueadores isolados devidamente ranqueados de forma decrescente.

5.2 Agregação das Listas Ranqueadas dos Descritores

O processo de combinação dos m melhores ranqueadores selecionados pela etapa de estimativa de eficácia consiste na agregação de cada ranqueador com os demais ranqueadores através de uma combinação par a par. Nessa fase é utilizado o primeiro método de agregação, como ilustrado pela Figura 23 e detalhado na descrição das etapas a seguir:

Figura 23 – Agregação de ranqueadores par-a-par.



Fonte: Autor

- **Etapa 4. Definição dos m melhores ranqueadores isolados:**

A partir dos resultados obtidos pela estimativa de eficácia definida, é aplicada uma filtragem dos primeiros m elementos, onde m é um parâmetro do método, criando assim uma lista ranqueada com m elementos;

Para a primeira camada de agregação é de extrema importância o foco em precisão (ao invés de revocação), uma vez que essas listas combinadas servirão de entrada para as etapas subsequentes. Dessa forma, foi inicialmente selecionada para a agregação nesta etapa o método RDPAC (PEDRONETTE; VALEM; LATECKI, 2021), uma vez que esse método tem atingido resultados positivos em cenários que exigem precisão nas primeiras posições das listas ranqueadas (VALEM; PEDRONETTE;

LATECKI, 2023a). Os métodos de agregação considerados estão presentes na biblioteca UDLF (VALEM; PEDRONETTE, 2017), que foi utilizada na metodologia de agregação multinível baseada em duas camadas de agregação.

- **Etapa 5. Agregação de ranqueadores em pares:**

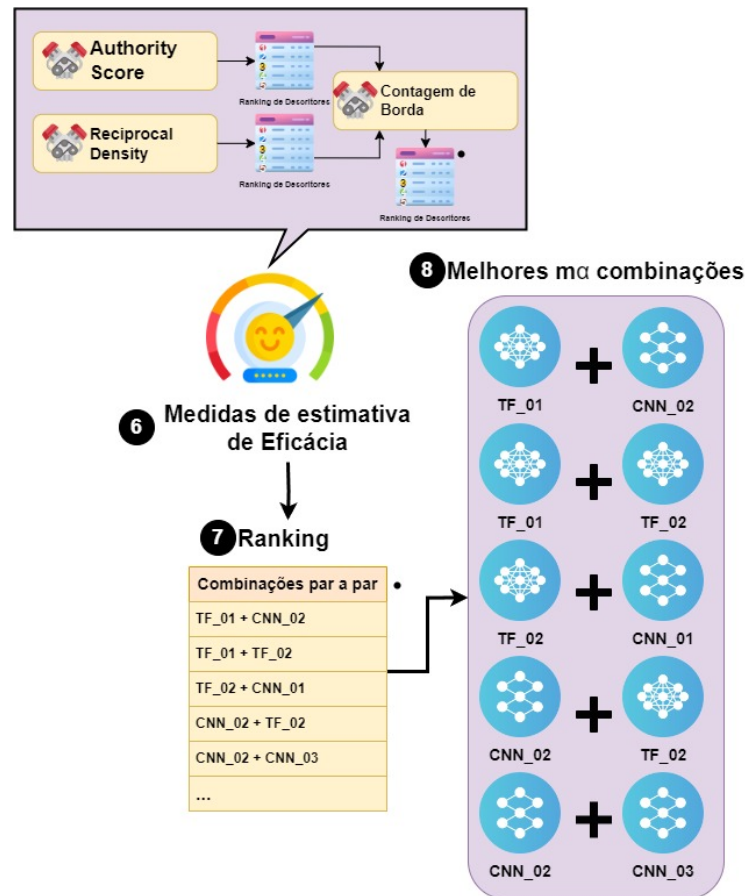
Utilizando-se dos ranqueadores filtrados na Etapa anterior, uma agregação par-a-par é executada gerando um produto cartesiano, cujo produto resultante é dado pela fórmula $\mathcal{R}_c = \frac{m!}{(m-p)! \times p!}$, em que p representa os número de elementos combinados simultaneamente (nessa caso pares, com $p = 2$).

Vale ressaltar a importância da seleção da metodologia adotada em cada etapa de fusão, uma vez que há numerosas abordagens de fusão disponíveis. Dentre as metodologias de fusão consideradas para a primeira camada, foi selecionado o método RDPAC, cujo foco principal se encontra em precisão, oferecendo uma aproximação eficiente de um processo de difusão baseado em dados de ranqueamento, combinando as vantagens dos métodos de difusão tradicionais com maior eficiência ao utilizar apenas as primeiras posições da lista ranqueada. (PEDRONETTE; VALEM; LATECKI, 2021).

5.3 Seleção de Melhores Combinações

O objetivo dessa etapa é selecionar as m_α melhores combinações baseadas em pares. Assim como o processo de estimativa de eficácia executado na seleção de descritores isolados, o objetivo desta etapa é obter uma lista com as melhores combinações geradas por meio da aplicação das medidas de estimativa de eficácia. A Figura 24 ilustra as etapas deste processo, discutido em detalhes a seguir:

Figura 24 – Seleção de ranqueadores combinados.



Fonte: Autor

- **Etapa 6. Estimativa de eficácia das combinações de ranqueadores:**

Nessa etapa são aplicadas novamente as medidas de estimativa de eficácia *Authority Score* e o *Reciprocal Score* e aplica-se a agregação dos resultados de ambas as métricas através da contagem de Borda, porém dessa vez as métricas são aplicadas nas listas de ranqueadores combinados, gerados pela primeira camada de combinação pareada a fim de obter uma nova lista ordenada;

- **Etapa 7. Lista ranqueada de ranqueadores combinados:**

Após a aplicação da estimativa de eficácia, o objeto de saída desse processamento é a lista de pares ranqueadores combinados ranqueados conforme a sua estimativa de eficácia;

- **Etapa 8. Definição dos m_α melhores ranqueadores combinados:**

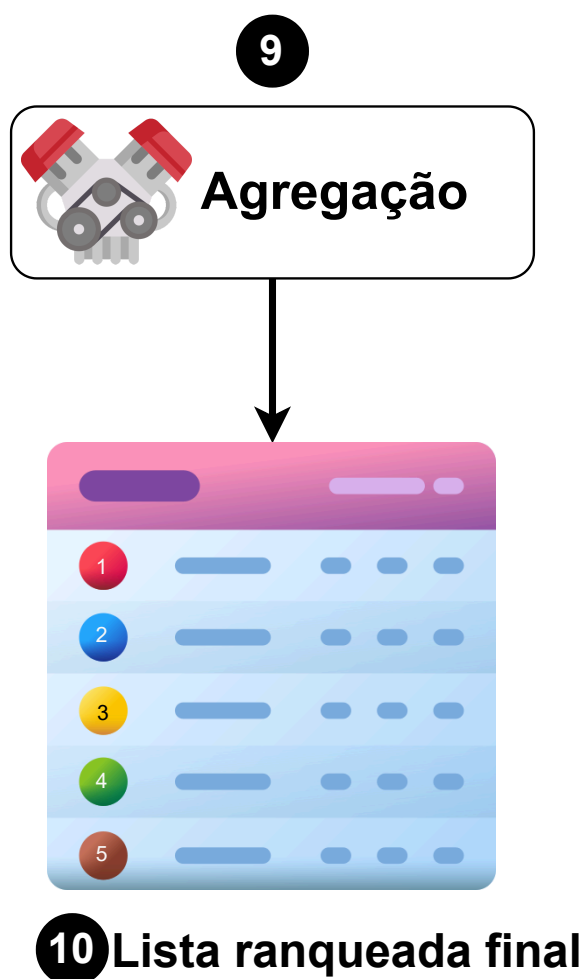
A partir dos resultados obtidos pela estimativa de eficácia aplicada, é aplicada uma filtragem dos primeiros m pares de ranqueadores combinados. Para essa etapa adota-se $m_\alpha = \alpha \times \mathcal{R}_c$, onde α equivale a um valor em um intervalo entre $]0, 1]$,

cuja aplicação determina a porcentagem de ranqueadores combinados resultantes da etapa descrita na Seção 5.2 que serão considerados para a agregação.

5.4 Agregação de Combinações

Essa etapa é a responsável pela fusão das m_α melhores combinações de ranqueadores obtidas. A Figura 25 ilustra as etapas envolvidas na etapa final de agregação, descritas a seguir:

Figura 25 – Ilustração do processo de agregação final.



Fonte:Autor

- **Etapa 9. Agregação das melhores combinações:**

Nessa etapa o processo de agregação é aplicado às m_α melhores listas ranqueadas obtidas através das combinações de ranqueadores em um procedimento final, gerando uma lista ranqueada final única. A natureza dessa agregação final é ser um complemento à primeira fusão, expandindo os resultados positivos obtidos na primeira etapa. Por se tratar de uma etapa de refinamento de resultados, esta etapa utiliza o método CPRR, cujo foco principal se encontra na revocação, buscando maximizar as informações de ranqueamento disponíveis utilizando operações de produto cartesiano sobre conjuntos de vizinhança, gerando novas relações de pares que revelam similaridades subjacentes (VALEM; PEDRONETTE; ALMEIDA, 2018).

- **Etapa 10. Lista ranqueada final:**

Após a aplicação da agregação, o resultado obtido é uma lista ranqueada dos elementos do *dataset* em ordem decrescente de similaridade como resultado.

6 Avaliação Experimental

Neste capítulo, será apresentada a avaliação experimental realizada para validação da abordagem proposta. Os *datasets* selecionados são discutidos na Seção 6.1, o protocolo experimental considerado é descrito na Seção 6.2 e os resultados obtidos são apresentados e discutidos na Seção 6.3. Considerando-se os resultados obtidos nos experimentos, a metodologia mostrou-se promissora, apresentando potencial para aumento das medidas de eficácia em tarefas de recuperação de imagens baseadas no conteúdo.

6.1 *Datasets*

A escolha dos *datasets* Cub200, descrito na Seção 6.1.3, Corel5K, descrito na Seção 6.1.2, e *Flowers-17*, visto na Seção 6.1.1, foi motivada por sua diversidade e representatividade para suas respectivas áreas de aplicação.

6.1.1 *Flowers-17*

Essa coleção de dados consiste em imagens de flores organizadas em 17 categorias distintas, cada uma das quais representa uma espécie de flor distinta, como ilustrado na Figura 26. Os algoritmos de aprendizado de máquina que usam essas imagens são principalmente usados para tarefas de reconhecimento, classificação e recuperação baseada em conteúdo de imagens. Cada categoria oferece uma variedade de fotos que mostram uma variedade de flores. Além de permitir que pesquisadores e cientistas criem e avaliem algoritmos de classificação de imagens, esse *dataset* é frequentemente usado para treinar e testar modelos de visão computacional (NILSBACK; ZISSERMAN, 2006; KUMAR; TRIPATHI; PANT, 2020).

Figura 26 – Amostra de elementos do conjunto Oxford *Flowers-17*.

Fonte:Autor (Baseado no conteúdo do *dataset*)

6.1.2 Corel5K

O *dataset* Corel5K é uma coleção de 5000 fotos que incluem uma variedade de categorias e temas. Essas imagens são muito utilizadas em pesquisas de visão computacional e processamento de imagens. O Corel5K oferece uma variedade de conteúdo visual categorizado em várias categorias, incluindo paisagens, animais, objetos e cenas urbanas, como pode ser observado na Figura 27. Esse *dataset* é útil para o desenvolvimento e teste de algoritmos de reconhecimento de padrões, classificação de imagens, recuperação de imagens baseada em conteúdo e outras aplicações de aprendizado de máquina, pois fornece uma variedade de exemplos visuais para treinamento e avaliação de modelos (CHENG et al., 2018; HUANG et al., 2010).

Figura 27 – Amostra de elementos do conjunto Corel5K.

Fonte:Autor (Baseado no conteúdo do *dataset*)

6.1.3 Cub200

O dataset Caltech-UCSD Birds-200 (CUB200) é composto por 200 espécies distintas de pássaros, cada espécie contendo cerca de 59 imagens, resultando em um total de 11.788 imagens. Projetado como uma referência de avaliação para algoritmos de reconhecimento visual de objetos, ele é especialmente relevante em cenários com uma grande diversidade de classes e variações sutis entre elas (WELINDER et al., 2010). Amplamente reconhecido para a classificação e detecção de pássaros em imagens, conforme pode ser observado na Figura 28.

Figura 28 – Amostra de elementos do conjunto Cub200.



Fonte: Autor (Baseado no conteúdo do *dataset*)

6.2 Protocolo Experimental, Parâmetros e Ranqueadores

Para avaliar a eficácia do método proposto nos *datasets* *Flowers-17*, *Corel5K* e *Cub200*, todas as imagens de cada *dataset* foram empregadas como imagens de consulta. Como medida de eficácia, considerou-se o MAP em todos os experimentos realizados.

A escolha de ranqueadores baseados em modelos de redes neurais convolucionais (CNNs) e *Transformers* foi motivada pelos resultados consistentemente superiores apresentados por esses métodos em comparação com ranqueadores clássicos baseados em cor, textura e forma. A eficácia desses ranqueadores baseados em modelos de aprendizado de máquina treinado via transferência de aprendizado têm sido evidenciada em diversas tarefas de processamento de imagens, demonstrando uma capacidade significativamente maior de capturar características complexas e abstratas. Ao adotar esses ranqueadores, visamos assegurar que o desempenho relativamente inferior dos ranqueadores clássicos não impacte negativamente nos resultados obtidos pela metodologia multinível, permitindo explorar plenamente o potencial das técnicas mais recentes na representação e extração de características de imagens.

Foram selecionadas modelos que tem atingido altos valores de eficácia limitado o número de modelos visando tornar mediano o custo computacional para a avaliação experimental: *Dual Path Networks* (DPN) (CHEN et al., 2017), *Residual Network* (RESNET) (HE et al., 2016), *Squeeze-and-Excitation Network* (SENET) (HU; SHEN; SUN, 2018), *Extreme Inception* (XCEPTION) (CHOLLET, 2017), *Swin Transformer* (SWINTF) (LIU et al., 2022), *Distillation with No Labels Version 2* (DINOv2) (OQUAB et al., 2023) e *Visual Transformer* (VIT-B16) (LIU et al., 2023b).

• Parâmetrização da metodologia

Para a realização dos experimentos, utilizamos os parâmetros $k = 100$, que representa a vizinhança considerada para as medidas de estimativa de eficácia, $m = 7$, ou seja, foram considerados todos os ranqueadores selecionados, $\alpha = 0,25$, e para o parâmetro L , responsável por determinar a profundidade a ser considerada do ranking, foi definida como padrão $L = 3500$ (exceto para o *dataset flowers*, no qual foi definido $L = 1360$ que é o tamanho do *dataset*). Os parâmetros foram definidos empiricamente por meio de experimentos e utilizados para todos os datasets. As metodologias de agregação possuem parâmetros internos próprios oriundos do *framework* UDLF, cuja parametrização foi mantida em algumas partes como o valor padrão do UDLF¹.

6.2.1 Resultados Isolados

Como primeira etapa da avaliação experimental, os ranqueadores foram avaliados individualmente. Para ilustrar os resultados das métricas de avaliação e estimativa de eficácia para cada descritor isolado, as Tabelas 3, 4 e 5 são apresentadas. Essas tabelas exibem os resultados individuais obtidos por cada descritor, apresentando sua eficácia (considerando a medida MAP), assim como os valores das estimativas de eficácia *Authority* e *Reciprocal* aplicadas e o ranking obtido pela aplicação da contagem de Borda às medidas de estimativa de eficácia. Pode-se observar que, para os 3 *datasets*, o primeiro descritor ranqueado pela contagem de Borda coincidiu com o maior valor de MAP (em negrito).

¹ RDPAC - valor inicial de k : $k_{inicial} = 1$, o valor de final de k : $k_{final} = k$, o valor de incremento para k : $k_{incremento} = 1$, dimensão da amostra: $RDPAC_L = \frac{L}{2}$, fator multiplicador de L : $Mult_L = 2$, Proporção de Variância: $P_{var} = 0.60$ e limiar de variância: $P_{lim} = 0.99$. CPRR - dimensão da amostra: $CPRR_L = L$, número de vizinhos: $CPRR_k = k$ e número de iterações: $CPRR_T = 2$.

Tabela 3 – Resultados isolados de cada descritor para o *dataset* Flowers-17 ordenado pelo valor da coluna Borda Rank.

Descritor	MAP	<i>Authority Score</i>	<i>Reciprocal Score</i>	Borda Rank
DINOv2	0,9859	0,77429	0,04481	1
SWINTF	0,9300	0,68245	0,04326	2
VIT-B16	0,8771	0,62453	0,04083	3
RESNET152	0,5183	0,48956	0,03532	4
XCEPTION	0,4731	0,49180	0,03528	5
DPN92	0,5093	0,47433	0,03445	6
SENET154	0,4316	0,46690	0,03441	7

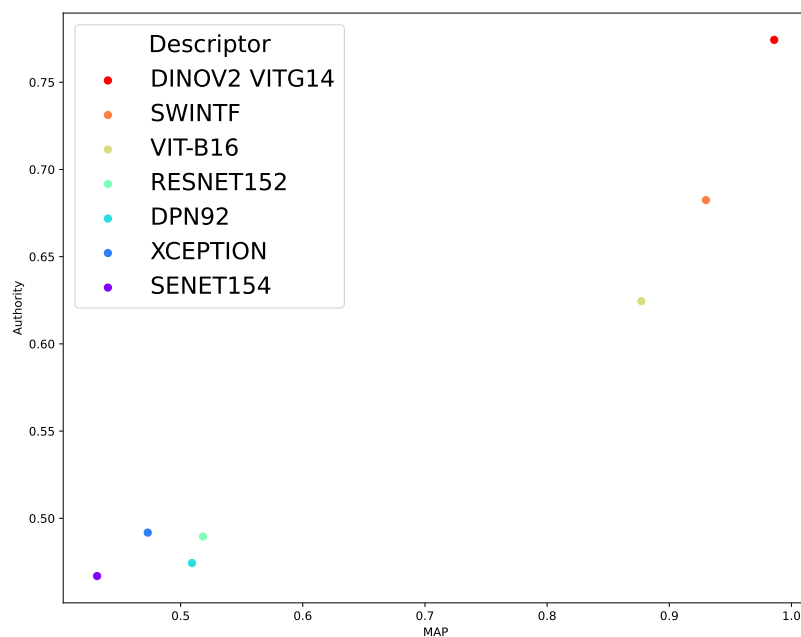
Tabela 4 – Resultados isolados de cada descritor para o *dataset* Corel5K ordenado pelo valor da coluna Borda Rank.

Descritor	MAP	<i>Authority Score</i>	<i>Reciprocal Score</i>	Borda Rank
DINOv2	0,7859	0,71541	0,04272	1
VIT-B16	0,7525	0,65472	0,04119	2
SWINTF	0,7434	0,65266	0,04174	3
DPN92	0,6516	0,57152	0,03822	4
SENET154	0,5699	0,55428	0,03831	5
RESNET152	0,6483	0,55950	0,03776	6
XCEPTION	0,5449	0,51590	0,03581	7

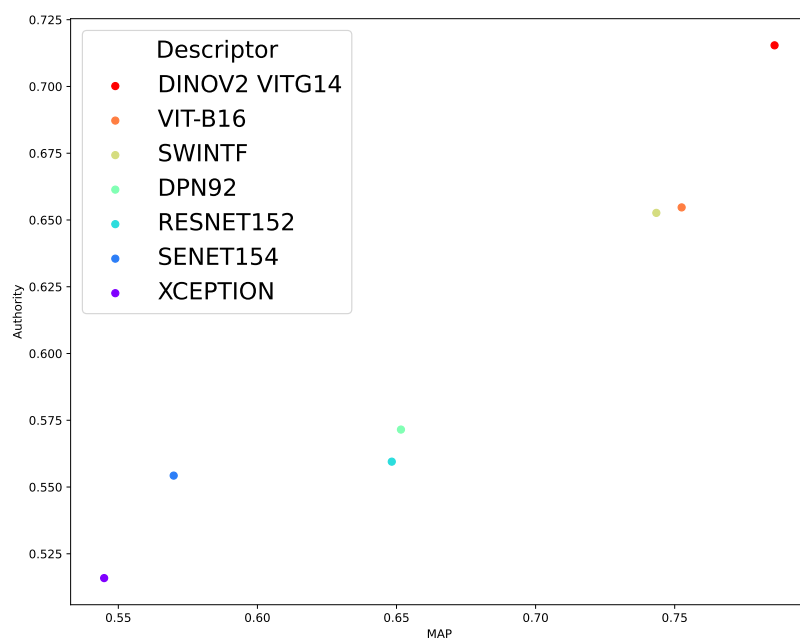
Tabela 5 – Resultados isolados de cada descritor para o *dataset* Cub200 ordenado pelo valor da coluna Borda Rank.

Descritor	MAP	<i>Authority Score</i>	<i>Reciprocal Score</i>	Borda Rank
DINOv2	0,6599	0,73211	0,04437	1
SWINTF	0,5840	0,61030	0,04079	2
VIT-B16	0,6082	0,58639	0,03900	3
SENET	0,1889	0,42350	0,03215	4
DPN92	0,2658	0,37702	0,03009	5
XCEPTION	0,2620	0,36840	0,02991	6
RESNET152	0,2324	0,31362	0,02704	7

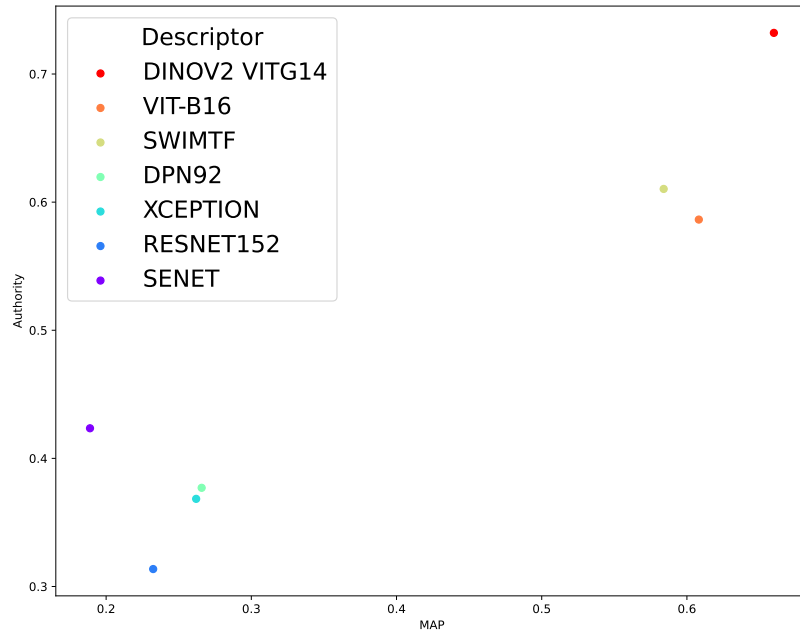
Analisando os resultados das tabelas, pode-se observar que os modelos baseados em *Transformers* tendem a apresentar um valor de MAP superior em comparação com as redes baseadas em convolução. A relação entre o resultado da estimativa de eficácia com a avaliação da média de eficácia (MAP) é apresentada nos gráficos das Figuras 29, 30 e 31.

Figura 29 – Gráfico de relação MAP e *Authority* para o dataset *flowers-17*.

Fonte:Autor

Figura 30 – Gráfico de relação MAP e *Authority* para o dataset Corel5K.

Fonte:Autor

Figura 31 – Gráfico de relação MAP e *Authority* para o *dataset* Cub200.

Fonte:Autor

Ao observar os gráficos apresentados nas Figuras 29, 30 e 31 fica evidente que a estimativa de eficácia está diretamente correlacionado ao MAP na maior parte dos ranqueadores. Isso indica que o processo de seleção baseado nas estimativas de eficácia é consistente.

6.3 Resultados de Combinação

Essa seção têm o objetivo de apresentar os resultados da avaliação experimental conduzida considerando combinações. Inicialmente serão apresentados os resultados de recuperação para os pares de descritores. Em seguida, serão apresentados os resultados obtidos para a abordagem UDLA completa, até a combinação final.

Os resultados obtidos para os pares de combinações para cada *dataset* obtidos pela primeira etapa de agregação (valores da etapa intermediária) do UDLA são apresentados nas Tabelas 6, 7, 8. Os resultados estão ordenados pela estimativa de eficácia dada pelo Borda Rank e o melhor valor de MAP obtido está destacado em negrito.

Tabela 6 – Resultados *dataset* Flowers-17 ordenados pelo Borda Rank

Borda rank	ranqueadores combinados	MAP
1	DINOv2 + SWINTF	0,9982
2	DINOv2 + ViT-B16	0,9948
3	DINOv2 + SENET154	0,9947
4	DINOv2 + XCEPTION	0,9956
5	DINOv2 + RESNET152	0,9970
6	SWINTF + ViT-B16	0,9907
7	SWINTF + DPN92	0,9846
8	SWINTF + XCEPTION	0,9855
9	DINOv2 + DPN92	0,9941
10	SWINTF + RESNET152	0,9876
11	SWINTF + SENET154	0,9810
12	ViT-B16 + XCEPTION	0,9718
13	ViT-B16 + RESNET152	0,9746
14	ViT-B16 + SENET154	0,9610
15	ViT-B16 + DPN92	0,9634
16	XCEPTION + DPN92	0,8038
17	RESNET152 + XCEPTION	0,7986
18	RESNET152 + DPN92	0,8051
19	RESNET152 + SENET154	0,7875
20	DPN92 + SENET154	0,7596
21	XCEPTION + SENET154	0,7723

Tabela 7 – Resultados *dataset* Corel5k ordenados pelo Borda Rank

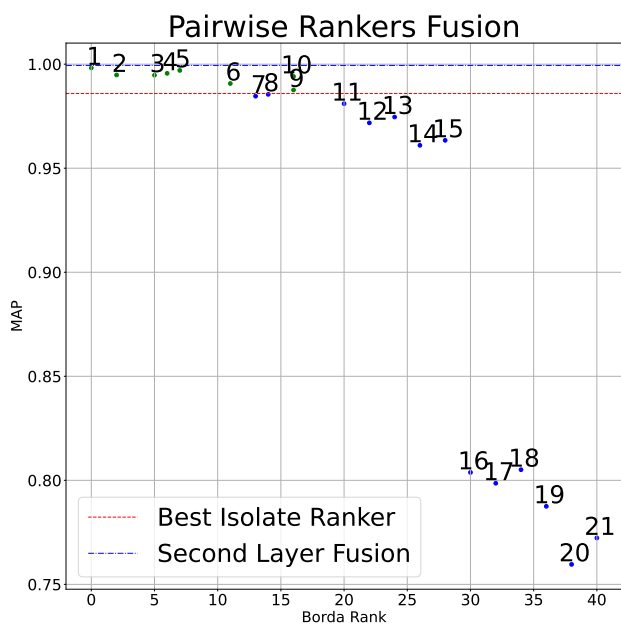
Borda rank	ranqueadores combinados	MAP
1	VIT-B16 + SWINTF	0,9385
2	SWINTF + DPN92	0,9414
3	SWINTF + RESNET152	0,9347
4	DINOV2 + RESNET152	0,9304
5	DINOV2 + DPN92	0,9308
6	DINOV2 + SWINTF	0,9324
7	VIT-B16 + RESNET152	0,9202
8	DINOV2 + VIT-B16	0,9248
9	VIT-B16 + DPN92	0,9212
10	SWINTF + XCEPTION	0,8908
11	SWINTF + SENET154	0,8886
12	DPN92 + RESNET152	0,8777
13	VIT-B16 + XCEPTION	0,8763
14	DINOV2 + XCEPTION	0,8777
15	DINOV2 + SENET154	0,8799
16	VIT-B16 + SENET154	0,8754
17	SENET154 + RESNET152	0,8737
18	DPN92 + XCEPTION	0,8343
19	RESNET152 + XCEPTION	0,8376
20	DPN92 + SENET154	0,8383
21	SENET154 + XCEPTION	0,8301

Tabela 8 – Resultados *dataset* Cub200 ordenados pelo Borda Rank

Borda rank	ranqueadores combinados	MAP
1	DINOv2 + SWINTF	0,7679
2	DINOv2 + ViT-B16	0,7736
3	SWINTF + ViT-B16	0,7675
4	DINOv2 + DPN92	0,6764
5	DINOv2 + RESNET152	0,6632
6	DINOv2 + XCEPTION	0,6737
7	SWINTF + XCEPTION	0,6652
8	SWINTF + DPN92	0,6601
9	SWINTF + RESNET152	0,6442
10	ViT-B16 + DPN92	0,6625
11	ViT-B16 + XCEPTION	0,6692
12	ViT-B16 + RESNET152	0,6487
13	DINOv2 + SENET	0,5854
14	SWINTF + SENET	0,5643
15	ViT-B16 + SENET	0,5692
16	DPN92 + XCEPTION	0,4745
17	DPN92 + RESNET152	0,4420
18	XCEPTION + RESNET152	0,4633
19	SENET + DPN92	0,3831
20	SENET + XCEPTION	0,4068
21	SENET + RESNET152	0,3691

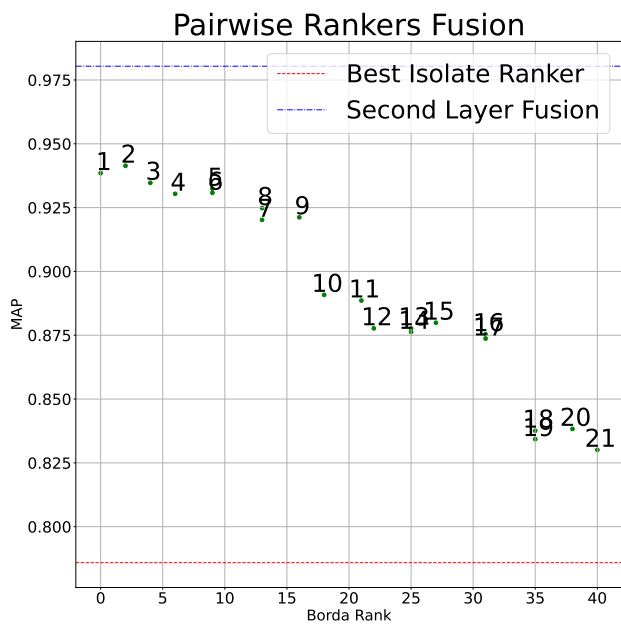
É possível observar nas Tabelas 6, 7 e 8 que as combinações que envolvem os melhores descritores isolados tendem a ficar em melhores posições quando ranqueadas. Essa situação fica evidente nos gráficos apresentados nas Figuras 32, 33 e 34. A linha vermelha pontilhada indica uma linha de base do valor do melhor descritor isolado e a linha tracejada em azul indica o resultado da aplicação da metodologia UDLA, permitindo uma visão comparativa entre os resultados do melhor isolado (Linha vermelha) e os m_α pares combinados. O valor de MAP da combinação dada pelo par de ranqueadores é representada no eixo y . A pontuação da estimativa de eficácia é representada no eixo x , dada pelos resultados do Método Borda aplicado para combinação do *Authority Score* e do *Reciprocal Score*.

Figura 32 – Gráfico de comparação de combinações e linha do melhor descritor isolado para o Flowers-17.

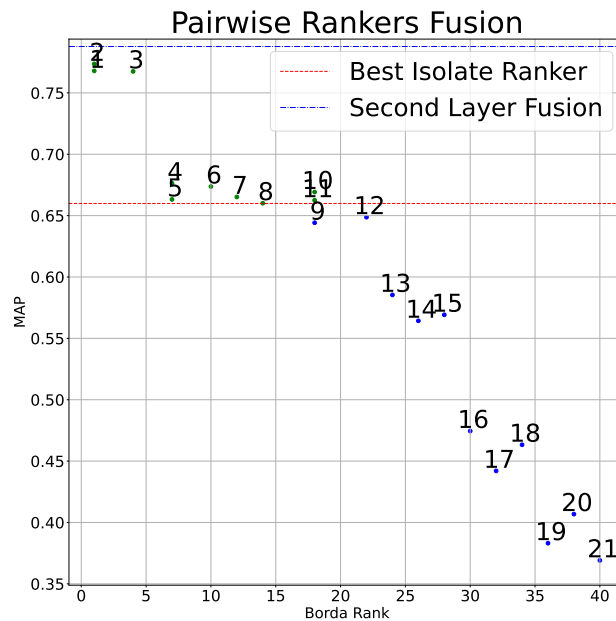


Fonte: Autor

Figura 33 – Gráfico de comparação de combinações e linha do melhor descritor isolado para o Corel5K.



Fonte: Autor

Figura 34 – Gráfico de comparação de combinações e linha do melhor descritor isolado para o Cub200.

Fonte:Autor

Ao analisar os gráficos é possível notar a correlação negativa entre o MAP dos pares e a pontuação de Método Borda. Esse comportamento indica que a adoção dessa metodologia tende a selecionar as combinações de pares mais eficazes, uma vez que a ordenação da contagem de Borda é no formato crescente, ou seja, quanto menor o seu valor, melhor será a posição do elemento avaliado. Além da avaliação dos descritores isolados e aos pares, consideramos a avaliação em relação ao melhor ranqueador isolado e uma abordagem de seleção e combinação recente da literatura.

Foram aplicados os mesmos parâmetros utilizados na metodologia UDLA na metodologia USRF (VALEM; PEDRONETTE, 2020c), cuja área de abordagem se assemelha à proposta pelo UDLA, permitindo uma comparação entre a metodologia UDLA e outra metodologia já consolidada na literatura. A Tabela 9 apresenta os valores comparativos utilizando o valor do MAP para cada *dataset*, destacando o melhor ranqueador isolado, a melhor combinação, o resultado obtido pela aplicação do UDLA e o resultado obtido pelo USRF.

Tabela 9 – Resultados obtidos com a aplicação do método UDLA para os *datasets* e a comparação com o método USRF.

<i>Dataset</i>	Melhor Ranqueador isolado	Melhor Combinação	UDLA	USRF
Flowers	0,9859	0,9982	0,9993	0,9972
Corel5k	0,7859	0,9414	0,9806	0,9773
Cub200	0,6599	0,7736	0,7878	0,7697

A partir destes resultados é observado que a metodologia UDLA apresentou resultados que se mostram relevantes, apresentando um ganho no resultado obtido com sua aplicação, demonstrando sua eficácia e seu potencial de aplicação.

6.4 Resultados Visuais

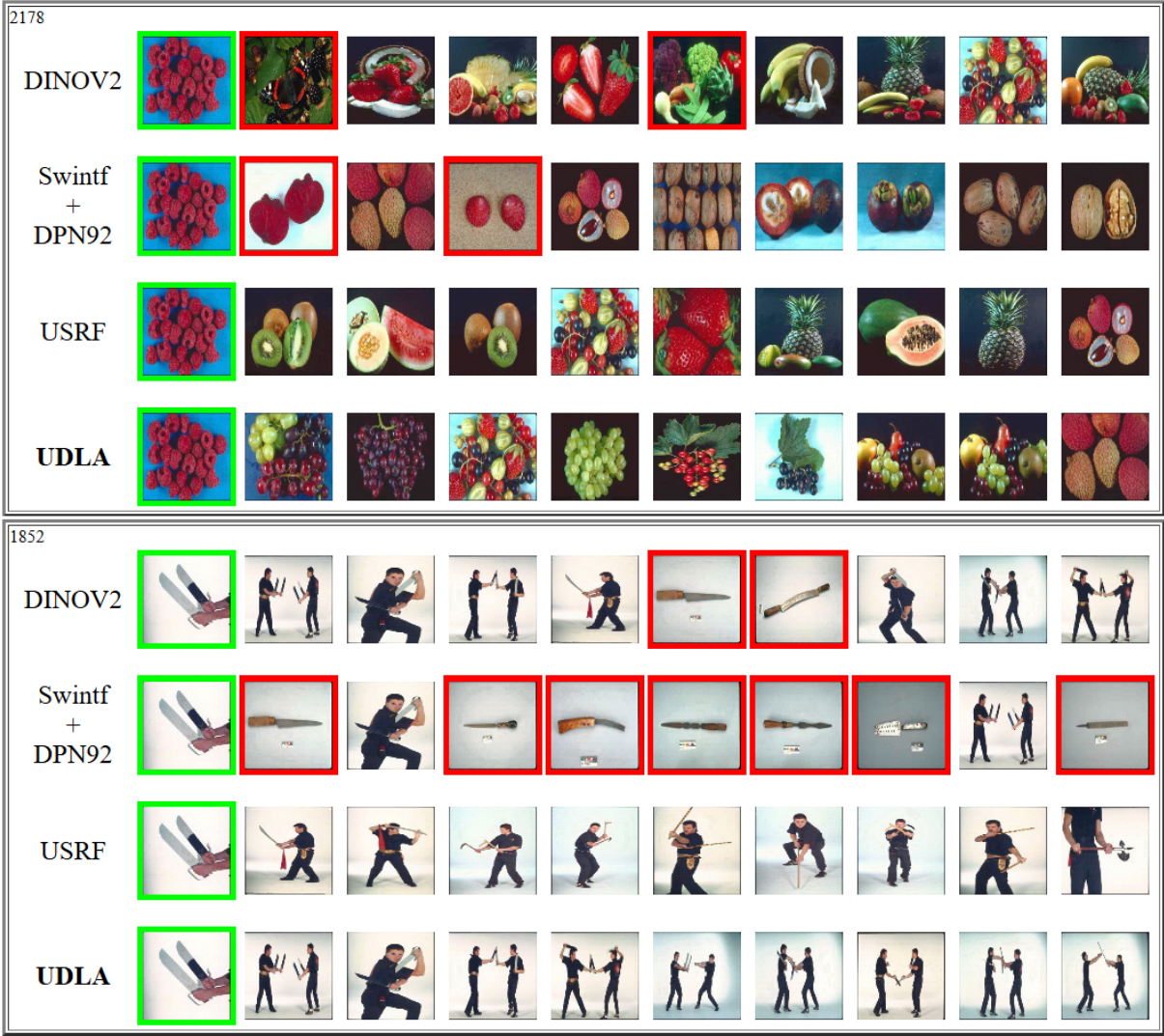
Para avaliação da relevância das imagens utilizou-se do arquivo de classes do *dataset*, por esse motivo as imagens que não pertencem a classe da imagem de consulta (irrelevantes) se destacam com a borda vermelha. As imagens foram selecionadas para ilustrar de forma mais clara os resultados das metodologias, sendo assim não possuem validade estatística.

É possível observar que existem ambiguidades do próprio *dataset*, a imagem 1200 na Figura 36 apresenta uma pessoa nadando com um cachorro, por exemplo, sua classe pelo *dataset* é definida como a classe das fotos de cães, em alguns resultados fica perceptível que os algoritmos detectam o nadador ou a piscina, o mesmo ocorre para a imagem 2157, apesar de ser uma fruta suas características acabam sendo mescladas as de grãos ou pedras. Tais características são características do próprio *dataset* e podem ocasionar a uma confusão de características para os algoritmos de recuperação.

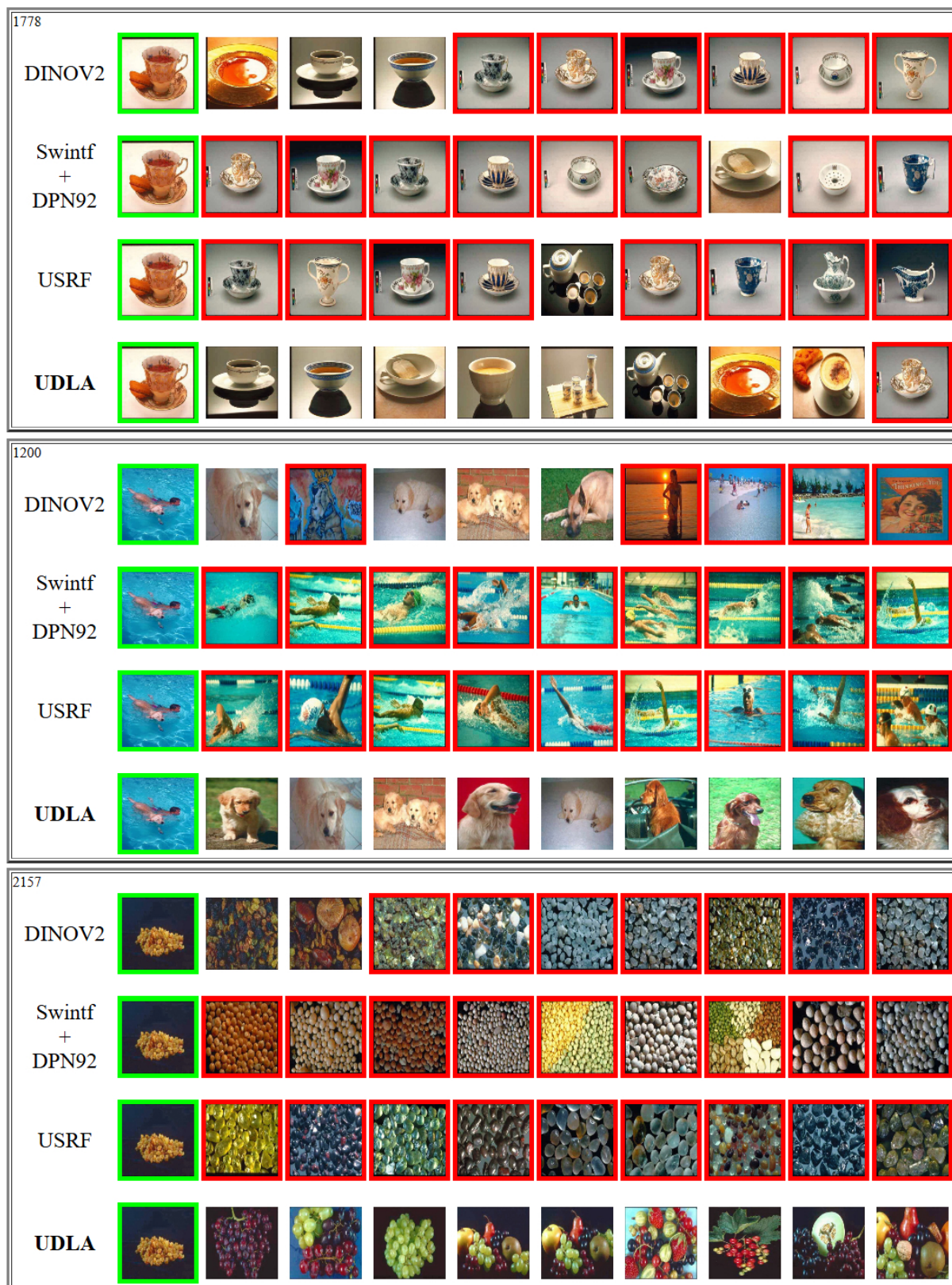
Essas ambiguidades no *dataset* ajudam a avaliar a robustez dos algoritmos de recuperação baseados em conteúdo. Quando os dados contêm exemplos desafiadores, como imagens com características mistas ou classificações ambíguas, é possível verificar até que ponto os algoritmos conseguem identificar corretamente as classes desejadas e lidar com sobreposição de características visuais. Isso fornece uma visão mais completa da capacidade dos algoritmos de diferenciar elementos semelhantes e interpretar corretamente o contexto das imagens, resultando em uma avaliação mais realista de seu desempenho em situações complexas.

As Figuras 35 e 36 apresentam de forma visual os resultados obtidos para o *dataset* Corel5k para o melhor ranqueador isolado, a melhor combinação de pares, para a aplicação da metodologia USRF (VALEM; PEDRONETTE, 2020c) e o resultado obtido pela aplicação do UDLA.

Figura 35 – Resultados visuais para as imagens 2178 e 1852 do *dataset* Corel5K.



Fonte:Autor

Figura 36 – Resultados visuais para as imagens 1178, 1200 e 2157 do *dataset* Corel5K.

Fonte:Autor

Os resultados obtidos indicam que o algoritmo UDLA, ao empregar uma abordagem

de agregação de ranqueadores em duas camadas, demonstra um bom desempenho quando comparado ao USRF. A primeira camada do UDLA, baseada na agregação par-a-par dos melhores ranqueadores, regida por uma metodologia baseada em precisão, permite explorar de forma abrangente o espaço de combinações possíveis. A segunda camada, por sua vez, refina essa busca ao selecionar as melhores combinações e agregá-las em uma lista final, regida por uma metodologia baseada em revocação.

Essa arquitetura em duas camadas, confere ao UDLA uma maior flexibilidade e capacidade de adaptação a diferentes tipos de coleções e consultas, demonstrando uma capacidade superior em capturar as nuances e complexidades inerentes ao problema de recuperação de informação.

- **Eficiência e Custo computacional**

A eficiência da metodologia se mostrou relevante para o cenário de recuperação de imagens não supervisionada com resultados promissores, entretanto durante as execuções e ressaltado o custo computacional da metodologia UDLA quando comparada ao USRF, essa diferença foi constatada empiricamente com as execuções de ambas as metodologias em testes, onde o USRF retornava o resultado em algumas horas no máximo, o UDLA exigia uma execução de alguns dias. Vale ressaltar que essa diferença de execução se deve a fatores de otimização que ainda podem ser aplicados à metodologia UDLA, que podem otimizar sua execução e reduzir a discrepância de tempo de execução entre as metodologias.

7 Conclusão

Este trabalho examinou técnicas de recuperação de imagens baseadas em conteúdo (CBIR) visando apresentar uma nova abordagem para agregação não supervisionada em duas camadas. A primeira camada inclui a combinação em pares dos m melhores ranqueadores selecionados por meio de estimativas de eficácia. A segunda camada agrega as melhores combinações obtidas na primeira etapa. Essa abordagem aproveita a complementariedade entre os ranqueadores selecionados para aumentar a eficácia da recuperação de imagens. Os resultados mostram que a metodologia é capaz de melhorar a qualidade dos resultados de recuperação de imagens, o que oferece uma perspectiva inovadora para futuros avanços no CBIR.

Este estudo destacou os desafios envolvidos na seleção de características visuais eficazes para a recuperação de imagens em cenários não supervisionados. O processo intrincado de alcançar resultados positivos em tarefas de agregação onde não há dados rotulados disponíveis foi destacado. A seleção e fusão adequadas de ranqueadores desempenham um papel crucial na recuperação de imagens baseada em conteúdo, influenciando diretamente a eficácia das listas de classificação geradas.

Outra contribuição da abordagem proposta consiste no uso de diferentes medidas de eficácia e diferentes métodos contextuais de agregação de listas. O método proposto, UDLA, utiliza medidas de estimativa de eficácia para a seleção não supervisionada de ranqueadores isolados e agregações par-a-par. O método inclui uma fusão na primeira camada focada na precisão e uma fusão na segunda camada focada na revocação, ambas baseadas em métodos de agregação de ranking contextual.

Uma extensa avaliação experimental, considerando características visuais baseadas em modelos de CNN e Transformer, indica a eficácia da abordagem proposta, alcançando resultados superiores à melhor característica isolada e a uma abordagem recente de fusão. Os experimentos revelaram que a combinação de precisão para a primeira camada e revocação para a segunda camada de agregação permite capturar uma gama mais ampla de resultados relevantes, obtendo um resultado que se trata de um refinamento das metodologias de fusão isoladas, enquanto inclui complementariedade nos rankings finais, uma vez que ao aumentar a precisão, o recall pode diminuir, e vice-versa. Um modelo pode ser configurado para ser mais preciso, arriscando-se a perder algumas detecções positivas, ou ser mais abrangente, mas com risco de incluir mais classificações incorretas.

O trabalho desenvolvido deu origem a um artigo científico aceito para publicação em conferência da área:

- Ademir Moreno Junior, Daniel Carlos Guimarães Pedronette. *Unsupervised Dual-Layer Aggregation for Feature Fusion on Image Retrieval Tasks* (JUNIOR, 2024), aceito para publicação na conferência *SIBGRAPI (Conference on Graphics, Patterns and Images)*, 2024.

Em trabalhos futuros, pretendemos avaliar os aspectos de eficiência do método e investigar estratégias de otimização de desempenho. Planejamos explorar o impacto da escalabilidade em grandes volumes de dados e examinar como o método se comporta em diferentes contextos de aplicação. Além disso, pretendemos analisar a sensibilidade do UDLA a diferentes configurações de parâmetros e sua adaptabilidade a novos cenários de recuperação de informação. O objetivo é garantir que a metodologia não apenas mantenha sua eficácia, mas também ofereça um desempenho robusto e eficiente em ambientes variados e desafiadores.

Referências

ABOUZAHIR, S.; SADIK, M.; SABIR, E. Bag-of-visual-words-augmented histogram of oriented gradients for efficient weed detection. *Biosystems Engineering*, v. 202, p. 179–194, 2021.

AFIFI, A. J.; ASHOUR, W. M. Image retrieval based on content using color feature. *International Scholarly Research Notices*, Hindawi, v. 2012, 2012.

AHMAD, S.; PAL, R.; GANIVADA, A. Rank level fusion of multimodal biometrics using genetic algorithm. *Multimedia Tools and Applications*, Springer, v. 81, n. 28, p. 40931–40958, 2022.

AKRITIDIS, L.; ALAMANIOTIS, M.; BOZANIS, P. Flagr: A flexible high-performance library for rank aggregation. *SoftwareX*, Elsevier, v. 21, p. 101319, 2023.

AKRITIDIS, L. et al. An unsupervised distance-based model for weighted rank aggregation with list pruning. *Expert Systems with Applications*, Elsevier, v. 202, p. 117435, 2022.

ALDAHDOOH, A. et al. Adversarial example detection for dnn models: A review and experimental comparison. *Artificial Intelligence Review*, Springer, v. 55, n. 6, p. 4403–4462, 2022.

ALKHAWLANI, M.; ELMOGY, M.; BAKRY, H. E. Text-based, content-based, and semantic-based image retrievals: a survey. *Int. J. Comput. Inf. Technol*, v. 4, n. 01, p. 58–66, 2015.

ANDREA, R.; IKHSAN, N.; SUDIRMAN, Z. Face recognition using histogram of oriented gradients with tensorflow in surveillance camera on raspberry pi. *International Journal of Information Engineering & Electronic Business*, v. 14, n. 1, 2022.

AWAD, A. I.; HASSABALLAH, M. Bag-of-visual-words for cattle identification from muzzle print images. *Applied Sciences*, MDPI, v. 9, n. 22, p. 4914, 2019.

AYALEW, A. M. et al. Detection and classification of covid-19 disease from x-ray images using convolutional neural networks and histogram of oriented gradients. *Biomedical Signal Processing and Control*, Elsevier, v. 74, p. 103530, 2022.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca*. [S.l.]: Bookman Editora, 2013.

BALCHANOWSKI, M.; BORYCZKA, U. Aggregation of rankings using metaheuristics in recommendation systems. *Electronics*, MDPI, v. 11, n. 3, p. 369, 2022.

BAO, T. Q. et al. Plant species identification from leaf patterns using histogram of oriented gradients feature space and convolution neural networks. *Journal of Information and Telecommunication*, Taylor & Francis, v. 4, n. 2, p. 140–150, 2020.

- BIBI, R. et al. Deep features optimization based on a transfer learning, genetic algorithm, and extreme learning machine for robust content-based image retrieval. *Plos one*, Public Library of Science San Francisco, CA USA, v. 17, n. 10, p. e0274764, 2022.
- CASTRO, L. N. D. *Fundamentals of natural computing: basic concepts, algorithms, and applications*. [S.l.]: CRC Press, 2006.
- CHEN, Y. et al. Dual path networks. *Advances in neural information processing systems*, v. 30, 2017.
- CHENG, Q. et al. A survey and analysis on automatic image annotation. *Pattern Recognition*, Elsevier, v. 79, p. 242–259, 2018.
- CHOLLET, F. Xception: Deep learning with depthwise separable convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 1251–1258.
- CORMACK, G. V.; CLARKE, C. L.; 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*. [S.l.: s.n.], 2009. p. 758–759.
- CRESTANI, F.; WU, S. Testing the cluster hypothesis in distributed information retrieval. *Information Processing & Management*, Elsevier, v. 42, n. 5, p. 1137–1150, 2006.
- DATTA, R. et al. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, ACM New York, NY, USA, v. 40, n. 2, p. 1–60, 2008.
- DOURADO, I. C.; PEDRONETTE, D. C. G.; TORRES, R. da S. Unsupervised graph-based rank aggregation for improved retrieval. *Information Processing & Management*, Elsevier, v. 56, n. 4, p. 1260–1279, 2019.
- DOURADO, I. C.; TORRES, R. d. S. Fusion vectors: Embedding graph fusions for efficient unsupervised rank aggregation. *arXiv preprint arXiv:1906.06011*, 2019.
- DUBEY, S. R. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE, v. 32, n. 5, p. 2687–2704, 2021.
- D’AMBROSIO, A. et al. Median constrained bucket order rank aggregation. *Computational Statistics*, Springer, v. 34, p. 787–802, 2019.
- ENG, S. et al. Facial expression recognition in jaffe and kdef datasets using histogram of oriented gradients and support vector machine. In: IOP PUBLISHING. *IOP Conference series: materials science and engineering*. [S.l.], 2019. v. 705, n. 1, p. 012031.
- FAGIN, R.; KUMAR, R.; SIVAKUMAR, D. Comparing top k lists. *SIAM Journal on discrete mathematics*, SIAM, v. 17, n. 1, p. 134–160, 2003.
- FAGIN, R.; KUMAR, R.; SIVAKUMAR, D. Efficient similarity search and classification via rank aggregation. In: *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*. [S.l.: s.n.], 2003. p. 301–312.

- FRÄNTI, P.; MARIESCU-ISTODOR, R. Soft precision and recall. *Pattern Recognition Letters*, Elsevier, v. 167, p. 115–121, 2023.
- GARCÍA-LAPRESTA, J. L.; MARTÍNEZ-PANERO, M.; MENESES, L. C. Defining the borda count in a linguistic decision making context. *Information Sciences*, Elsevier, v. 179, n. 14, p. 2309–2316, 2009.
- GEETHA, V. et al. Efficient hybrid multi-level matching with diverse set of features for image retrieval. *Soft Computing*, Springer, v. 24, p. 12267–12288, 2020.
- GIONIS, A. et al. Algorithms for discovering bucket orders from data. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2006. p. 561–566.
- HAMEED, I. M.; ABDULHUSSAIN, S. H.; MAHMMOD, B. M. Content-based image retrieval: A review of recent trends. *Cogent Engineering*, Taylor & Francis, v. 8, n. 1, p. 1927469, 2021.
- HE, F.; LIU, T.; TAO, D. Why resnet works? residuals generalize. *IEEE Transactions on Neural Networks and Learning Systems*, v. 31, n. 12, p. 5349–5362, 2020.
- HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778.
- HU, J.; SHEN, L.; SUN, G. Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 7132–7141.
- HUANG, Y. et al. Image retrieval via probabilistic hypergraph ranking. In: IEEE. *2010 IEEE computer society conference on computer vision and pattern recognition*. [S.l.], 2010. p. 3376–3383.
- IONESCU, B. et al. *Fusion in Computer Vision: Understanding Complex Visual Content*. [S.l.]: Springer Science & Business Media, 2014.
- JABNOUN, J. et al. An image retrieval system using deep learning to extract high-level features. In: SPRINGER. *Advances in Computational Collective Intelligence: 14th International Conference, ICCCI 2022, Hammamet, Tunisia, September 28–30, 2022, Proceedings*. [S.l.], 2022. p. 167–179.
- JARDINE, N.; RIJSBERGEN, C. J. van. The use of hierarchic clustering in information retrieval. *Information storage and retrieval*, Elsevier, v. 7, n. 5, p. 217–240, 1971.
- JUNIOR, D. C. G. P. A. M. Unsupervised dual-layer aggregation for feature fusion on image retrieval tasks. In: *37th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI*. [S.l.]: IEEE, 2024.
- KAUR, N.; NAZIR, N. et al. A review of local binary pattern based texture feature extraction. In: IEEE. *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. [S.l.], 2021. p. 1–4.
- KLEMENTIEV, A.; ROTH, D.; SMALL, K. An unsupervised learning algorithm for rank aggregation. In: SPRINGER. *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17–21, 2007. Proceedings 18*. [S.l.], 2007. p. 616–623.

- KOBAYASHI, M.; TAKEDA, K. Information retrieval on the web. *ACM computing surveys (CSUR)*, ACM New York, NY, USA, v. 32, n. 2, p. 144–173, 2000.
- KOVÁCS, Z. L. *Redes neurais artificiais*. [S.l.]: Editora Livraria da Física, 2002.
- KUMAR, S. et al. A hierarchical approach based cbir scheme using shape, texture, and color for accelerating retrieval process. *Journal of King Saud University-Computer and Information Sciences*, Elsevier, p. 101609, 2023.
- KUMAR, V.; TRIPATHI, V.; PANT, B. Content based fine-grained image retrieval using convolutional neural network. In: IEEE. *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*. [S.l.], 2020. p. 1120–1125.
- KUNCHEVA, L. I. *Combining pattern classifiers: methods and algorithms*. [S.l.]: John Wiley & Sons, 2014.
- KURLAND, O. The cluster hypothesis in information retrieval. In: SPRINGER. *Advances in Information Retrieval: 36th European Conference on IR Research, ECIR 2014, Amsterdam, The Netherlands, April 13-16, 2014. Proceedings 36*. [S.l.], 2014. p. 823–826.
- LI, J. et al. Subspace-based multi-view fusion for instance-level image retrieval. *The Visual Computer*, Springer, v. 37, p. 619–633, 2021.
- LI, Z. et al. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, IEEE, 2021.
- LIN, H. et al. An ensemble matchers based rank aggregation method for taxonomy matching. In: SPRINGER. *Web Technologies and Applications: 17th Asia-Pacific Web Conference, APWeb 2015, Guangzhou, China, September 18-20, 2015, Proceedings 17*. [S.l.], 2015. p. 190–202.
- LIU, T. *Learning to rank for information retrieval*. *Found Trends Inf Retr.* 2009; 3 (3): 225–331.
- LIU, Y. et al. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, p. 1–21, 2023.
- LIU, Y. et al. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, IEEE, 2023.
- LIU, Y.-T. et al. Supervised rank aggregation. In: *Proceedings of the 16th international conference on World Wide Web*. [S.l.: s.n.], 2007. p. 481–490.
- LIU, Z. et al. Swin transformer v2: Scaling up capacity and resolution. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2022. p. 12009–12019.
- LIU, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF international conference on computer vision*. [S.l.: s.n.], 2021. p. 10012–10022.
- MA, W. et al. Adaptive multi-feature fusion via cross-entropy normalization for effective image retrieval. *Information Processing & Management*, Elsevier, v. 60, n. 1, p. 103119, 2023.

- MICROSOFT, B. I. C. *cachorro sobre uma cama rasgando um travesseiro com plano de fundo de um quarto*. 2023. Microsoft Bing Image Creator. Gerada em 30 de Abril de 2023. Disponível em: <<https://www.bing.com/images/create/cachorro-sobre-uma-cama-rasgando-um-travesseiro-co/64714f74cc8f4d3cbdb7fdc266cf3270?FORM=GUH2CR>>.
- MYERS, L.; SIROIS, M. J. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences*, Wiley Online Library, v. 12, 2004.
- NILSBACK, M.-E.; ZISSERMAN, A. A visual vocabulary for flower classification. In: *IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2006. v. 2, p. 1447–1454.
- NIXON, M.; AGUADO, A. *Feature extraction and image processing for computer vision*. [S.l.]: Academic press, 2019.
- OKADA, C. Y.; PEDRONETTE, D. C. G.; TORRES, R. da S. Unsupervised distance learning by rank correlation measures for image retrieval. In: *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. [S.l.: s.n.], 2015. p. 331–338.
- OQUAB, M. et al. Dinov2: Learning robust visual features without supervision. *arXiv e-prints*, p. arXiv–2304, 2023.
- PATHAK, D.; RAJU, U. Content-based image retrieval using feature-fusion of groupnormalized-inception-darknet-53 features and handcraft features. *Optik*, Elsevier, v. 246, p. 167754, 2021.
- PEDRONETTE, D. C. G.; PENATTI, O. A.; TORRES, R. d. S. Unsupervised manifold learning using reciprocal knn graphs in image re-ranking and rank aggregation tasks. *Image and Vision Computing*, Elsevier, v. 32, n. 2, p. 120–130, 2014.
- PEDRONETTE, D. C. G.; TORRES, R. d. S. Image re-ranking and rank aggregation based on similarity of ranked lists. *Pattern Recognition*, Elsevier, v. 46, n. 8, p. 2350–2360, 2013.
- PEDRONETTE, D. C. G.; TORRES, R. d. S. Unsupervised effectiveness estimation for image retrieval using reciprocal rank information. In: IEEE. *2015 28th SIBGRAP Conference on Graphics, Patterns and Images*. [S.l.], 2015. p. 321–328.
- PEDRONETTE, D. C. G.; TORRES, R. da S. Shape retrieval using contour features and distance optimization. In: *VISAPP (2)*. [S.l.: s.n.], 2010. p. 197–202.
- PEDRONETTE, D. C. G. et al. Multimedia retrieval through unsupervised hypergraph-based manifold ranking. *IEEE Transactions on Image Processing*, IEEE, v. 28, n. 12, p. 5824–5838, 2019.
- PEDRONETTE, D. C. G.; VALEM, L. P.; LATECKI, L. J. Efficient rank-based diffusion process with assured convergence. *Journal of Imaging*, Mdpi, v. 7, n. 3, p. 49, 2021.
- PEDRONETTE, D. C. G. et al. Semi-supervised and active learning through manifold reciprocal knn graph for image retrieval. *Neurocomputing*, Elsevier, v. 340, p. 19–31, 2019.
- PIETIKÄINEN, M. Local binary patterns. *Scholarpedia*, v. 5, n. 3, p. 9775, 2010.

- PINTO, V. B. Indexação morfossemântica de imagens no contexto da saúde visando à recuperação de informações. *História, Ciências, Saúde-Manguinhos*, SciELO Brasil, v. 15, p. 313–330, 2008.
- PIRAS, L.; GIACINTO, G. Information fusion in content based image retrieval: A comprehensive overview. *Information Fusion*, Elsevier, v. 37, p. 50–60, 2017.
- QI, J.; YANG, C.-H. H.; TEJEDOR, J. Submodular rank aggregation on score-based permutations for distributed automatic speech recognition. In: IEEE. *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2020. p. 3517–3521.
- RENDA, M. E.; STRACCIA, U. Web metasearch: rank vs. score based rank aggregation methods. In: *Proceedings of the 2003 ACM symposium on Applied computing*. [S.l.: s.n.], 2003. p. 841–846.
- RUBLEE, E. et al. Orb: An efficient alternative to sift or surf. In: IEEE. *2011 International conference on computer vision*. [S.l.], 2011. p. 2564–2571.
- SAIN, A. et al. Cross-modal hierarchical modelling for fine-grained sketch based image retrieval. *arXiv preprint arXiv:2007.15103*, 2020.
- SALMAN, R.; ALZAATREH, A.; SULIEMAN, H. The stability of different aggregation techniques in ensemble feature selection. *Journal of Big Data*, SpringerOpen, v. 9, n. 1, p. 1–23, 2022.
- SARICA, A.; QUATTRONE, A.; QUATTRONE, A. Introducing the rank-biased overlap as similarity measure for feature importance in explainable machine learning: A case study on parkinson’s disease. In: MAHMUD, M. et al. (Ed.). *Brain Informatics*. Cham: Springer International Publishing, 2022. p. 129–139. ISBN 978-3-031-15037-1.
- SCHALEKAMP, F.; ZUYLEN, A. v. Rank aggregation: Together we’re strong. In: SIAM. *2009 Proceedings of the Eleventh Workshop on Algorithm Engineering and Experiments (ALENEX)*. [S.l.], 2009. p. 38–51.
- SCULLEY, D. Rank aggregation for similar items. In: SIAM. *Proceedings of the 2007 SIAM international conference on data mining*. [S.l.], 2007. p. 587–592.
- SHAW, J. A.; FOX, E. A. Combination of multiple searches. *NIST SPECIAL PUBLICATION SP*, NATIONAL INSTITUTE OF STANDARDS & TECHNOLOGY, p. 105–105, 1995.
- ŠKRLJ, B.; REPAR, A.; POLLAK, S. Rank-based keyword extraction via unsupervised learning and meta vertex aggregation. In: SPRINGER. *Statistical Language and Speech Processing: 7th International Conference, SLSP 2019, Ljubljana, Slovenia, October 14–16, 2019, Proceedings 7*. [S.l.], 2019. p. 311–323.
- SMEULDERS, A. W. et al. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 22, n. 12, p. 1349–1380, 2000.
- SNOEK, C. G.; WORRING, M.; SMEULDERS, A. W. Early versus late fusion in semantic video analysis. In: *Proceedings of the 13th annual ACM international conference on Multimedia*. [S.l.: s.n.], 2005. p. 399–402.

- STUART, R.; PETER, N. *Artificial Intelligence A Modern Approach Third Edition*. [S.l.]: PEARSON, 2010.
- SWAIN, M. J.; BALLARD, D. H. Color indexing. *International journal of computer vision*, Springer, v. 7, n. 1, p. 11–32, 1991.
- TEJASHWINI, S.; ARADHANA, D. Revolutionizing sentiment classification: A deep learning approach using self-attention based encoding–decoding transformers with feature fusion. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 125, p. 106730, 2023.
- TORRES, R. D. S.; FALCÃO, A. X. Content-based image retrieval: Theory and applications. *Revista de Informática Teórica e Aplicada*, v. 13, p. 161–185, 2006.
- TYAGI, V.; RETRIEVAL, C.-B. I. *Ideas, Influences, and Current Trends*. [S.l.]: London: Springer, 2018.
- VALEM, L. P. Combinação seletiva não supervisionada de listas ranqueadas aplicada à busca de imagens pelo conteúdo. Universidade Estadual Paulista (Unesp), 2019.
- VALEM, L. P.; PEDRONETTE, D. C. G. Unsupervised selective rank fusion on content-based image retrieval. In: SBC. *Anais Estendidos do XXXII Conference on Graphics, Patterns and Images*. [S.l.], 2019. p. 63–69.
- VALEM, L. P.; PEDRONETTE, D. C. G. Graph-based selective rank fusion for unsupervised image retrieval. *Pattern Recognition Letters*, Elsevier, v. 135, p. 82–89, 2020.
- VALEM, L. P.; PEDRONETTE, D. C. G. Unsupervised selective rank fusion for image retrieval tasks. *Neurocomputing*, Elsevier, v. 377, p. 182–199, 2020.
- VALEM, L. P.; PEDRONETTE, D. C. G. Unsupervised selective rank fusion for image retrieval tasks. *Neurocomputing*, v. 377, p. 182 – 199, 2020. ISSN 0925-2312. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0925231219313335>>.
- VALEM, L. P.; PEDRONETTE, D. C. G.; ALMEIDA, J. Unsupervised similarity learning through cartesian product of ranking references. *Pattern Recognition Letters*, Elsevier, v. 114, p. 41–52, 2018.
- VALEM, L. P.; PEDRONETTE, D. C. G.; LATECKI, L. J. Graph convolutional networks based on manifold learning for semi-supervised image classification. *Computer Vision and Image Understanding*, Elsevier, v. 227, p. 103618, 2023.
- VALEM, L. P.; PEDRONETTE, D. C. G.; LATECKI, L. J. Rank flow embedding for unsupervised and semi-supervised manifold learning. *IEEE Transactions on Image Processing*, IEEE, 2023.
- VALEM, L. P.; PEDRONETTE, D. C. G. a. An unsupervised distance learning framework for multimedia retrieval. In: *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. New York, NY, USA: ACM, 2017. (ICMR '17), p. 107–111. ISBN 978-1-4503-4701-3. Disponível em: <<http://doi.acm.org/10.1145/3078971.3079017>>.
- WANG, S. et al. From coarse to fine: multi-level feature fusion network for fine-grained image retrieval. *Multimedia Systems*, Springer, v. 28, n. 4, p. 1515–1528, 2022.

- WANG, Y. et al. Hybrid cnn-transformer features for visual place recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE, 2022.
- WELINDER, P. et al. Caltech-ucsd birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.
- WISLER, C. The spearman correlation formula. *Science*, American Association for the Advancement of Science, v. 22, n. 558, p. 309–311, 1905.
- WU, Z.; YU, J. A multi-level descriptor using ultra-deep feature for image retrieval. *Multimedia Tools and Applications*, Springer, v. 78, p. 25655–25672, 2019.
- XIAO, Y. et al. Graph-based rank aggregation method for high-dimensional and partial rankings. *Journal of the Operational Research Society*, Taylor & Francis, v. 72, n. 1, p. 227–236, 2021.
- YAO, W. et al. Early and late fusion of multiple modalities in sentinel imagery and social media retrieval. In: SPRINGER. *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII*. [S.l.], 2021. p. 591–606.
- ZHANG, K. et al. Content-based image retrieval with a convolutional siamese neural network: Distinguishing lung cancer and tuberculosis in ct images. *Computers in biology and medicine*, Elsevier, v. 140, p. 105096, 2022.
- ZHANG, Z. et al. Effective image retrieval via multilinear multi-index fusion. *IEEE Transactions on Multimedia*, IEEE, v. 21, n. 11, p. 2878–2890, 2019.