Unsupervised Dual-Layer Aggregation for Feature Fusion on Image Retrieval Tasks

Ademir Moreno Junior
Department of Statistics
Applied Mathematics and Computing
São Paulo State University (UNESP)
Rio Claro, Brazil
ademir.junior@unesp.br

Daniel Carlos Guimarães Pedronette
Department of Statistics
Applied Mathematics and Computing
São Paulo State University (UNESP)
Rio Claro, Brazil
daniel.pedronette@unesp.br

***Abstract*—The revolutionary advances in image representation have led to impressive progress in many image understanding-related tasks, primarily supported by Convolutional Neural Networks (CNN) and, more recently, by Transformer models. Despite such advances, assessing the similarity among images for retrieval in unsupervised scenarios remains a challenging task, mostly grounded on traditional pairwise measures, such as the Euclidean distance. The scenario is even more challenging when different visual features are available, requiring the selection and fusion of features without any label information. In this paper, we propose an Unsupervised Dual-Layer Aggregation (UDLA) method, based on contextual similarity approaches for selecting and fusing CNN and Transformer-based visual features trained through transfer learning. In the first layer, the selected features are fused in pairs focused on precision. A sub-set of pairs is selected for a second layer aggregation focused on recall. An experimental evaluation conducted in different public datasets showed the effectiveness of the proposed approach, which achieved results significantly superior to the best-isolated feature and also superior to a recent fusion approach considered as baseline.**

I. Introduction

Information retrieval (IR) is a domain in constant metamorphosis, driven by the continuous advancement of computational capabilities, especially with the integration of machine learning-based methodologies aiming for efficient and effective retrieval of the vast and expansive repository of available content [1]. Especially in the visual domain, a huge growth in image collections was observed, mainly triggered by a crescent storage of information in visual and multimedia formats. In this scenario, Content-Based Image Retrieval (CBIR) systems have emerged as a relevant solution in IR methodologies, responding to the growing demand and the development of computational technologies [2], [3].

During decades of evolution and development of CBIR systems, there has been a notable proliferation of a diversified set of feature extraction methodologies. In many cases, the application of an isolated extraction technique proves insufficient to capture the multiple aspects of an image, resulting in low-efficiency ranked lists. Faced with this issue, a promising solution is to explore different aggregation techniques to generate a single, final ranked list. Aggregation techniques can be implemented at different stages of the fusion process, the most common being during the initial stages, known as early fusion, and after feature processing, known as late fusion [4].

Notably, aggregation methodologies exhibit substantial gains compared to isolated rankers, as observed in various approaches proposed by different authors [5]–[8]. Initially exploited for aggregating different handcrafted features based on visual properties given by shape [9], color [10], and texture [11], [12], aggregation methodologies remain significantly, despite the impressive progress in the deep learning techniques.

In fact, the advent of deep features based on Convolutional Neural Networks (CNNs) [13], and Visual Transformers [14], [15] lead to impressive results. Trained through transfer learning on large-scale datasets (often Imagenet), such features establish the basis for state-of-the-art unsupervised image retrieval. Nevertheless, considering the large number of available approaches, how to exploit the complementary of distinct models remains a challenging research question, especially in unsupervised scenarios. In the absence of labeled data, selecting the results from a wide variety of deep-based models is an intricate and complex task.

This paper addresses this challenge by proposing a novel rank-based fusion for unsupervised image retrieval tasks. The Unsupervised Dual-Layer Aggregation (UDLA) method exploits effectiveness estimation measures to select a subset of features. The combinations of pairs of features are fused in a first-layer contextual rank aggregation focused on precision. The results of pairs are re-selected by effectiveness estimation measures for a second layer contextual rank aggregation focused on recall. The proposed method was experimentally evaluated on three public and distinct datasets and various features based on CNNs and Transformers models. The retrieval results are superior to the best-isolated feature and recent rank aggregation approach.

Other works already have proposed multi-level ranked list aggregation methodologies, which treat the levels differently for each methodology. Such methodologies aim to use aggregation methodologies at more than one execution level. In [6], a hierarchical multi-level framework is proposed, where each feature group is considered in a single hierarchy, with the main concern being the proper selection of the order of visual feature retrieval. In [7], a methodology based on three levels responsible for specific actions is proposed. The color/contour level is responsible for extracting contour and color characteristics and concatenating them into a single Vector of Locally Aggregated Descriptors (VLAD) descriptor. The partial level deals with the intermediate convolution layers, which tend to capture information from object parts. However, to the best of our knowledge, this work is the first to use contextual rank aggregation methods [16], [17] in a dual-layer approach for fusing CNNs and Transformers models in image retrieval.

The presentation structure of this paper will follow the following organization: Section II discusses the formal definition used in this work; in Section III, the proposed two-layer aggregation model is presented; Section IV discusses the experimental protocol conducted for assessing the effectiveness of the proposed model, as well as the comparison with the best-isolated feature and other aggregation models.

II. Rank Model and Problem Definition

This section formally defines the ranking model used throughout this paper. Let $\mathcal{C} = \{x_1, x_2, x_3, \ldots, x_n\}$ be an image collection, where n denotes the size of the collection. Consider a content-based retrieval task where, given a query image, it returns a list of images from the collection $\mathcal{C}$.

Formally, given a query image x_q, a ranker denoted by R_j computes a ranked list $\tau_q = (x_1, x_2, x_3, \ldots, x_k)$ in response to the query. The ranked list τ_q can be defined by the permutation of the k neighbors in $\mathcal{N}(q)$, which contains the k most similar images to image x_q in the collection $\mathcal{C}$. The permutation τ_q is a bijection from the set $\mathcal{N}(q)$ to the set $[k] = \{1, 2, 3, \ldots, k\}$. The notation $\tau_q(i)$ denotes the position of image x_i in the ranked list τ_q.

A ranker R_j can be defined based on different features and distance functions, such that each combination can produce distinct ranked lists. In this sense, a ranker can be seen as a descriptor [18], which is formally defined as a tuple (ϵ, ρ), where $\epsilon : \mathcal{C} \rightarrow \mathbb{R}^d$ is a function that extracts a feature vector $\mathbf{x}_i$ from an image $x_i \in \mathcal{C}$, and ρ: $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a distance function that computes the distance between two images according to their corresponding feature vectors.

Considering the diversity of rankers available, we can defined $\mathcal{R} = \{R_1, R_2, R_3, \ldots, R_r\}$ as a the set of rankers. In this paper, the distance function is given by the Euclidean distance for all rankers. The function ϵ for each ranker is given by different CNNs and Transformer-based models trained through transfer learning on ImageNet [19] dataset.

Considering a given ranker R_j and taking each image $x_i \in \mathcal{C}$ as a query, we can obtain a set of ranked lists $\mathcal{T}_j$. Such set of ranked lists can be exploited by contextual approaches for ranking and retrieval tasks [16], [17]. The research challenge addressed in this paper consists of how to combine different sets of ranked list $\mathcal{T}_j$ defined by each ranker $R_j \in \mathcal{R}$ in order to compute an aggregated set of ranked lists $\mathcal{T}_a$. In this sense, the proposed approach UDLA can be defined as a function f_a, such that:

$$\mathcal{T}_a = f_a(\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_r). \tag{1}$$

The objective consists in exploiting contextual and complementary information in the different set of ranked lists to compute a more effective set $\mathcal{T}_a$.

III. Unsupervised dual-layer aggregation

This section discusses the proposed Unsupervised Dual-Layer Aggregation (UDLA) method, which addresses the complex tasks of selecting and combining a set of visual features in a fully unsupervised setting.

A. Overview of UDLA

The absence of labels in unsupervised scenarios leads to challenging questions in selection and fusion retrieval approaches: (*i*) How to determine (or estimate) the effectiveness of retrieval results of a given ranker? and (*ii*) How to define the best combination of rankers? The proposed UDLA approach addresses such questions by exploiting effectiveness estimation measures in two layers of aggregation.

Figure 1 illustrates the overall organization of UDLA, highlighting the main stages (Steps 1-10) of the method. The expected input of the method consists of the retrieval results of the set of rankers $\mathcal{R}$ (Step 1). The retrieval results are given by the set of ranked lists: for each ranker R_j a set of ranked lists $\mathcal{T}_j$ is defined.

In Step (2), an unsupervised effectiveness estimation is computed for each ranker, assigning a score that allows ranking the isolated rankers according to the estimated quality of retrieval results (Step 3). From the list of rankers, the top-m rankers are selected and become the input for the first aggregation phase (Step 4), where m is a parameter of the method. Once properly defining the subset of rankers can directly affect the results, we devise an approach for turning the method robust to this definition. The subset is combined pair-by-pair such that the best pairs can be selected in the next layer.

The pairs are aggregated employing a contextual rank aggregation method focused on precision [16] (Step 5). Subsequently, effectiveness estimation measures are employed again, now to estimate the quality of retrieval results of aggregated pairs (Step 6) and to define a rank of pairs (Step 7). According to this ranking, the best pairs are selected (Step 8). To conclude the process, a final contextual aggregation focused on recall [17] is performed (Step 9) giving rise to the final retrieval results (Step 10).

In the following sections, the main steps of the proposed method are discussed and defined in detail.

B. Selection of Isolated Rankers

An effective methodology for selecting rankers is a crucial and challenging task in the unsupervised aggregation process, given that inadequate selection of rankers can significantly compromise the effectiveness of the aggregated results. In this paper, we exploit unsupervised measures used to estimate the effectiveness of ranked lists. Given a query image, the ranked lists obtained for it, and the ranked lists of the top-k neighborhood, such measures assign a real value that aims to estimate the qualify of retrieval results. The measures are mainly based on the cluster hypothesis [20], which states that closely associated elements tend to be relevant to the same requests. In this sense, the measures analyze the reciprocal references among elements at the top positions of ranked lists to define the estimated value. Two measures were considered:

- **Authority Measure:** The authority measure [21] is an effectiveness estimation measure based on the density of the k-neighborhood graph in a ranked list of an image x_q. The effectiveness estimation provided by the authority measure is based on the density of the neighborhood graph of an image x_q.
- **Reciprocal Density:** The reciprocal neighborhood density [22] also exploits the neighborhood information to

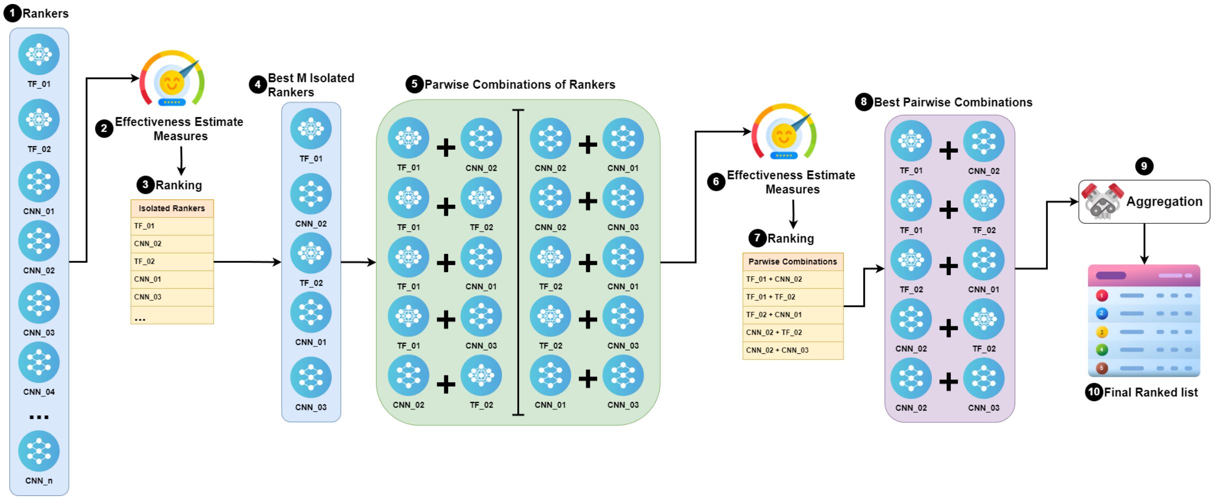

Fig. 1: General view and organization of the proposed Unsupervised Dual-Layer Aggregation (UDLA) method.

check the reciprocal references among elements in top-k positions of ranked lists. However, while the authority scores consider that all references have the same importance, the reciprocal density assigns higher weights to references at the top position of ranked lists.

Results of studies [23] indicate that different effectiveness estimation measures can be combined leading to more accurate estimates. In this work, we combine both measures through a traditional rank aggregation approach, the Borda [24] method. The Borda method is based on the positions of rankings. In this way, a list of rankers is obtained for each measure (Authority and Reciprocal Density). In the following, the position of each ranker in both lists is summed to compute a Borda score. The aggregated ranked list is obtained by sorting the rankers in crescent order of Borda score. The resulting ordered list is then used to select the top-m rankers in the ranking, which are used to select a subset of isolated rankers $\mathcal{R}_m \subset \mathcal{R}$.

C. Pairwise Aggregation

Given the selected subset of isolated rankers, all the possible combinations of pairs are computed. Formally, the set of combined pairs $\mathcal{R}_c$ can be defined by a Cartesian product of the set of isolated rankers, as $\mathcal{R}_c = \mathcal{R}_m \times \mathcal{R}_m$, composed by only non-repeating pairs. The number of computed aggregated pairs is given by $r_p = \binom{m}{p}$, where p denotes the size of tuples combined, and therefore $p = 2$ for pairs. The first aggregation layer aims to bring relevant results to top-positions of ranked lists, such that they can be exploited by the next layer. Therefore, an aggregation method focused on precision should be used for the fusion of pairs. Based on this assumption, we employed the Rank Diffusion with Assured Convergence (RDPAC) [16], a contextual re-ranking and rank aggregation method. Recent studies [25] indicated that RDPAC [16] can achieve high-accuracy results on classification tasks, which are grounded on the precision of top-ranking positions. The method takes as input the set of ranked lists for both rankers which defines the pair and produces as output one combined set of ranked lists for each pair in $\mathcal{R}_c$.

D. Selection of aggregated pairs

Once the pairs are combined, not all the pairs are used for the final aggregation. In this way, similar to the process of selecting individual rankers, this step selects a subset of pairs. The same approach defined in Section III-B is used for pairs selection, employing the unsupervised effectiveness estimation measures. A subset of the top m_c combined pairs is selected. The value of m_c is given by $m_c = r_p \times \alpha$, where r_p denotes the number of pairs and α is a parameter defined in a range between $(0, 1]$. The definition of α is associated with thresholding for pairs selection, which defines the proportion of combined rankers belonging to the set of rankers generated by the first aggregation layer that will be considered for the final fusion.

E. Final Aggregation

In this step, a final aggregation is conducted to fuse the results of selected pairs. Since it is expected that the retrieval results of pairs were improved especially in precision (at the top position of ranked lists), this step aims to exploit it to improve recall. In this way, we employ the contextual aggregation approach given by the Cartesian Product of Ranking References (CPRR) [17]. This method aims to maximize the available similarity information contained in top-ranking positions. By using Cartesian product operations on neighborhood sets, these operations create new similarity relationships [17].

IV. Experimental Evaluation

This section describes the experimental evaluation conducted to assess the effectiveness of the proposed method. In Section IV-A, the experimental protocol applied, along with the datasets and the selected rankers/ visual features are discussed. In Section IV-B, the results obtained from the pairwise aggregation of the first layer of the methodology are analyzed. In Section IV-C, the final results for each dataset are presented, as well as the comparison with a recent baseline. In Section IV-D, the visual results are presented.

A. Experimental Protocol, Datasets and Features

For the evaluation of the proposed method, the adopted protocol employed all the images contained in each dataset as a query. The effectivenes measure considered is the traditional Mean Average Precision (MAP). In the following, we discuss the datasets, visual features, and implementation details regarding the experimental evaluation:

- **Datasets:** The experimental analysis considered three datasets containing from 1,360 to 11,788 images and with different complexities. For experimentation, Flowers17 [26], Corel5k [27], and Cub200 [28] were used. The MAP was calculated considering each image in the dataset as a query.
- **Visual Features:** Six visual features were selected for the experiments. Traditional CNNs and recent Transformer-based models were considered, namely: Dual Path Networks (DPN) [29], Residual Network (RESNET) [30], Squeeze-and-Excitation Network (SENET) [31], Extreme Inception (XCEPTION) [32], Swin Transformer (SWINTF) [33], and Visual Transformer (VIT-B16) [34].

Tables I, II e III present the retrieval results achieved by the visual features in each dataset. The results of effectiveness estimation measures are also reported. We can observe that the values of these measures are highly correlated with the MAP scores.

TABLE I: Retrieval results for isolated features (MAP) and effectiveness estimation measures (Authority and Reciprocal) for Flowers17 dataset.

Ranking	**Descriptor**	**Original MAP**	**Authority**	**Reciprocal**
1	SWINTF	**0.9300**	0.89558	0.06195
2	VIT-B16	0.8771	0.81210	0.05857
3	RESNET152	0.5183	0.49102	0.04432
4	XCEPTION	0.4731	0.48584	0.04398
5	DPN92	0.5093	0.47779	0.04350
6	SENET154	0.4316	0.45529	0.04283

TABLE II: Retrieval results for isolated features (MAP) and effectiveness estimation measures (Authority and Reciprocal) for Corel5k dataset.

Ranking	**Descriptor**	**Original MAP**	**Authority**	**Reciprocal**
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

TABLE III: Retrieval results for isolated features (MAP) and effectiveness estimation measures (Authority and Reciprocal) for Cub200 dataset.

Ranking	**Descriptor**	**Original MAP**	**Authority**	**Reciprocal**
1	SWINTF	0.5840	0.61887	0.05149
2	VIT-B16	0.6082	0.60159	0.04960
3	SENET	0.1889	0.40342	0.03933
4	DPN92	0.2658	0.37041	0.03743
5	XCEPTION	0.2620	0.36310	0.03723
6	RESNET152	0.2324	0.30944	0.03372

- **Implementation details:** For the fusion methodologies, the unsupervised distance learning framework (UDLF) [35], which includes several pre-implemented fusion methods, was used. Regarding the parameters, we used $k = 100$ and $L = 3,500$ for contextual rank aggregation methods and effectiveness measures (except for Flowers17 dataset which used $L = 1,360$). The parameter L defines the length of ranked lists. For the first layer aggregation method RDPAC [16], we used $k = 15$, the default parameter of UDLF [35] framework. For UDLA we used $\alpha = 0.25$, $m = 6$ for all experiments in all datasets.

In order to compare with other approaches, the first natural baseline consisted of surpassing the performance of the best individual ranker, followed by exceeding the performance of the best pairwise fusion of descriptors. With these goals achieved, the next step was to compare the obtained results with another recent selection and aggregation method, such as the Unsupervised Selective Rank Fusion (USRF) [8].

We can observe that the values of these measures are highly correlated with the MAP scores.

B. Results of Pairwise Combination

This section aims to evaluate the capacity of the proposed method to achieve effectiveness gains when combining the pairs. In addition, we aim to evaluate the capacity of selecting the effective pairs. The results are presented in Tables IV,V, and VI.

TABLE IV: Retrieval results for pairwise combination of features (MAP) and effectiveness estimation measures of pairs (Authority and Reciprocal) for Flowers17 dataset.

Ranking	**Combined Descriptors**	**MAP**	**Authority**	**Reciprocal**
1	SWINTF + VIT-B16	**0.9950**	0.94854	0.06256
2	SWINTF + XCEPTION	0.9878	0.94329	0.06253
3	VIT-B16 + XCEPTION	0.9808	0.93727	0.06238
4	SWINTF + SENET154	0.9845	0.93658	0.06230
5	VIT-B16 + RESNET152	0.9776	0.93565	0.06236
6	SWINTF + RESNET152	0.9898	0.93581	0.06223
7	SWINTF + DPN92	0.9875	0.93525	0.06216
8	VIT-B16 + DPN92	0.9652	0.91916	0.06188
9	VIT-B16 + SENET154	0.9660	0.91369	0.06156
10	XCEPTION + DPN92	0.8109	0.78787	0.05726
11	RESNET152 + DPN92	0.8090	0.78403	0.05709
12	RESNET152 + XCEPTION	0.8029	0.78720	0.05704
13	RESNET152 + SENET154	0.7927	0.78040	0.05679
14	DPN92 + SENET154	0.7648	0.74915	0.05569
15	XCEPTION + SENET154	0.7771	0.74621	0.05554

TABLE V: Retrieval results for pairwise combination of features (MAP) and effectiveness estimation measures of pairs (Authority and Reciprocal) for Corel5k dataset.

Ranking	**Combined Descriptors**	**MAP**	**Authority**	**Reciprocal**
1	SWINTF + VITB16	**0.9380**	0.88978	0.04851
2	SWINTF + DPN92	0.9373	0.88063	0.04828
3	SWINTF + RESNET152	0.9318	0.88131	0.04823
4	VITB16 + RESNET152	0.9198	0.86719	0.04782
5	VITB16 + DPN92	0.9182	0.86229	0.04769
6	SWINTF + XCEPTION	0.8904	0.84446	0.04697
7	SWINTF + SENET154	0.8894	0.83439	0.04661
8	VITB16 + XCEPTION	0.8769	0.82830	0.04630
9	DPN92 + RESNET152	0.8771	0.82468	0.04645
10	VITB16 + SENET154	0.8782	0.82207	0.04614
11	SENET154 + RESNET152	0.8750	0.81150	0.04606
12	DPN92 + SENET154	0.8420	0.79953	0.04536
13	DPN92 + XCEPTION	0.8358	0.79896	0.04536
14	RESNET152 + XCEPTION	0.8374	0.79655	0.04533
15	SENET154 + XCEPTION	0.8347	0.79120	0.04520

Figures 2, 3, and 4 also present the results in a different perspective. Each pair is represented by a point in the graph. The number represents the rank position in the ranking of pairs. The values are also shown in Tables IV,V, and VI. We can observe in Figures 2, 3, 4 that most of the pairs achieved results superior to the best-isolated features (red line). The first pairs are significantly superior. We can also observe that the final aggregation result surpass the best isolated feature and all the pairs.

C. Final Aggregation Results and Comparisons

Table VII presents the results obtained for each dataset, comparing the performance of the best-isolated ranker, the

TABLE VI: Retrieval results for pairwise combination of features (MAP) and effectiveness estimation measures of pairs (Authority and Reciprocal) for Cub200 dataset.

Ranking	Combined Descriptors	MAP	Authority	Reciprocal
1	SWINTFF + VIT-B16	**0.7675**	0.70782	0.04275
2	SWINTFF + XCEPTION	0.6652	0.64816	0.04016
3	SWINTFF + DPN92	0.6601	0.64562	0.04005
4	SWINTFF + RESNET152	0.6442	0.64356	0.03976
5	VIT-B16 + DPN92	0.6625	0.63610	0.03991
6	VIT-B16 + XCEPTION	0.6692	0.63840	0.03989
7	VIT-B16 + RESNET152	0.6487	0.62442	0.03938
8	SWINTFF + SENET	0.5643	0.58493	0.03773
9	VIT-B16 + SENET	0.5692	0.57486	0.03753
10	DPN92 + XCEPTION	0.4745	0.54971	0.03646
11	DPN92 + RESNET152	0.4420	0.53411	0.03574
12	XCEPTION + RESNET152	0.4633	0.53153	0.03563
13	SENET + DPN92	0.3831	0.50009	0.03476
14	SENET + XCEPTION	0.4068	0.49711	0.03457
15	SENET + RESNET152	0.3691	0.48574	0.03401

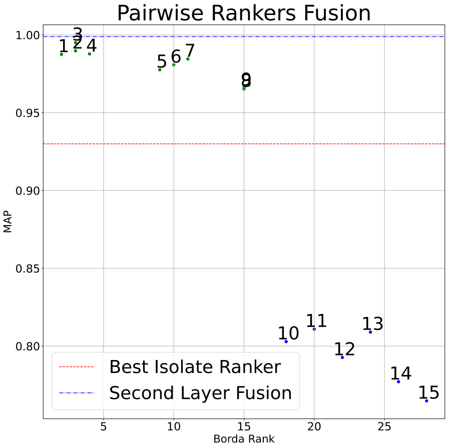

Fig. 2: Analysis of Pairwise Combination compared to the Best Isolated feature (red) and proposed approach UDLA (blue) for Flowers17 dataset.

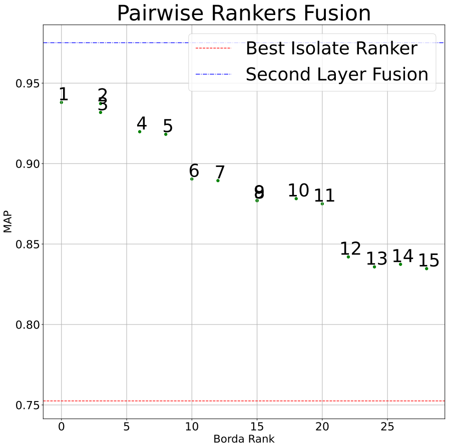

Fig. 3: Analysis of Pairwise Combination compared to the Best Isolated feature (red) and proposed approach UDLA (blue) for Corel5k dataset.

best-combined pair, the USRF approach [8], and the proposed UDLA approach.

The proposed UDLA approach achieved very significant effectiveness gains over the best-isolated ranker. For Corel5k dataset, we can observe an impressive improvement from 75.25% to 97.51%. The proposed approach also achieved results superior to the best pair and the best USRF [8] combination in all datasets.

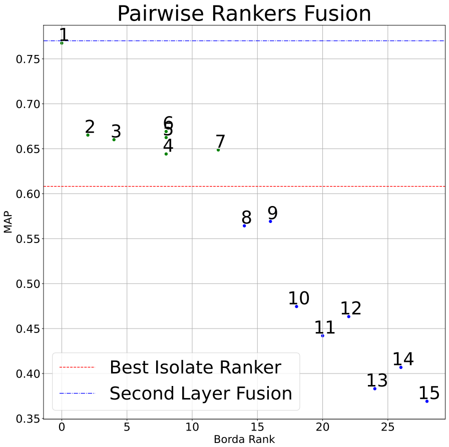

Fig. 4: Analysis of Pairwise Combination compared to the Best Isolated feature (red) and proposed approach UDLA (blue) for Cub200 dataset

TABLE VII: Final Aggregation Results of UDLA compared to other approaches: *Best Isolated Descriptor*, *Best Pair*, and *Best USRF* [8] result, considering MAP scores.

Dataset	Best Isolated	Best Pair	Best USRF [8]	Final UDLA
Flowers17	0.9300	0.9950	0.9973	**0.9989**
Corel5k	0.7525	0.9380	0.9679	**0.9751**
Cub200	0.6082	0.7675	0.7472	**0.7701**

D. Visual Results

Figures 5 and 6 present the visual results of UDLA in comparison with the two best-isolated descriptors and the best-combined pair. The query image is represented in green borders (at left). Each line represents a ranked list obtained by each approach. The non-relevant results are illustrated in red borders. From this visualization, it is possible to observe the increase in effectiveness in the results achieved through the proposed methodology.

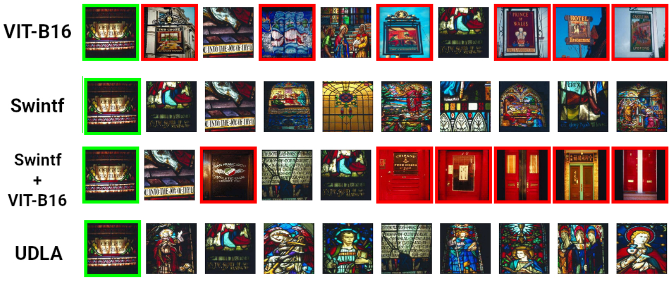

Fig. 5: Visual examples for Corel5k dataset image 53.

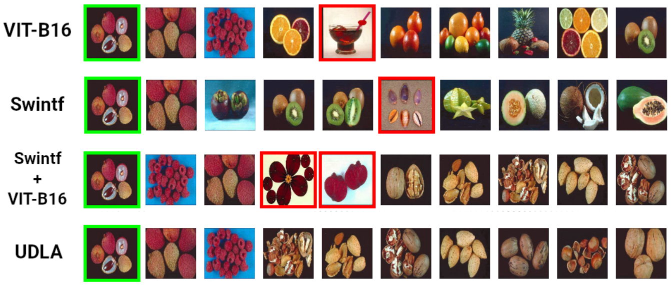

Fig. 6: Visual examples for Corel5k dataset image 2109.

V. Conclusion

This study highlighted the challenges involved in selecting effective visual features for image retrieval in unsupervised scenarios. It also emphasized the intricate process of achieving positive results in aggregation tasks where no labeled data is available. Properly selecting and fusing rankers play a crucial role in content-based image retrieval, as they directly influence the effectiveness of the generated ranked lists. In this paper, we proposed a dual-layer aggregation approach for image retrieval. The proposed UDLA method exploits effectiveness estimation measures for an unsupervised selection of isolated rankers and pairwise aggregations. The method includes a first-layer fusion focused on precision and a second-layer fusion focused on recall, both based on contextual rank aggregation methods. An extensive experimental evaluation considering visual features based on CNN and Transformer-based models indicates the effectiveness of the proposed approach, achieving results superior to the best-isolated feature and a recent fusion approach. In future work, we intend to evaluate the efficiency aspects of the method and investigate performance optimization strategies.

Acknowledgment

The authors are grateful to São Paulo Research Foundation - FAPESP (grants #2022/07349-8 and #2018/15597-6), Brazilian National Council for Scientific and Technological Development - CNPq (grants #313193/2023-1, and #422667/2021-8), and Petrobras (grant #2023/00095-3) for financial support. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

References

[1] R. Salman, A. Alzaatreh, and H. Sulieman, "The stability of different aggregation techniques in ensemble feature selection," *Journal of Big Data*, vol. 9, no. 1, pp. 1–23, 2022.

[2] L. P. Valem and D. C. G. Pedronette, "Unsupervised selective rank fusion on content-based image retrieval," in *Anais Estendidos do XXXII Conference on Graphics, Patterns and Images*. SBC, 2019, pp. 63–69.

[3] L. P. Valem, "Combinação seletiva não supervisionada de listas ranqueadas aplicada à busca de imagens pelo conteúdo," 2019.

[4] W. Yao, A. Moumtzidou, C. O. Dumitru, S. Andreadis, I. Gialampoukidis, S. Vrochidis, M. Datcu, and I. Kompatsiaris, "Early and late fusion of multiple modalities in sentinel imagery and social media retrieval," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII*. Springer, 2021, pp. 591–606.

[5] F. Schalekamp and A. v. Zuylen, "Rank aggregation: Together we're strong," in *2009 Proceedings of the Eleventh Workshop on Algorithm Engineering and Experiments (ALENEX)*. SIAM, 2009, pp. 38–51.

[6] S. Kumar, A. K. Pal, N. Varish, I. Nurhidayat, S. M. Eldin, and S. K. Sahoo, "A hierarchical approach based cbir scheme using shape, texture, and color for accelerating retrieval process," *Journal of King Saud University-Computer and Information Sciences*, p. 101609, 2023.

[7] Z. Wu and J. Yu, "A multi-level descriptor using ultra-deep feature for image retrieval," *Multimedia Tools and Applications*, vol. 78, pp. 25 655–25 672, 2019.

[8] L. P. Valem and D. C. G. Pedronette, "Unsupervised selective rank fusion for image retrieval tasks," *Neurocomputing*, vol. 377, pp. 182 – 199, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231219313335

[9] D. C. G. Pedronette and R. da Silva Torres, "Shape retrieval using contour features and distance optimization." in *VISAPP (2)*, 2010, pp. 197–202.

[10] M. J. Swain and D. H. Ballard, "Color indexing," *International journal of computer vision*, vol. 7, no. 1, pp. 11–32, 1991.

[11] M. Pietikäinen, "Local binary patterns," *Scholarpedia*, vol. 5, no. 3, p. 9775, 2010.

[12] N. Kaur, N. Nazir *et al.*, "A review of local binary pattern based texture feature extraction," in *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2021, pp. 1–4.

[13] R. Stuart and N. Peter, "Artificial intelligence a modern approach third edition," 2010.

[14] S. Tejashwini and D. Aradhana, "Revolutionizing sentiment classification: A deep learning approach using self-attention based encoding–decoding transformers with feature fusion," *Engineering Applications of Artificial Intelligence*, vol. 125, p. 106730, 2023.

[15] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2023.

[16] D. C. Guimarães Pedronette, L. Pascotti Valem, and L. J. Latecki, "Efficient rank-based diffusion process with assured convergence," *Journal of Imaging*, vol. 7, no. 3, p. 49, 2021.

[17] L. P. Valem, D. C. G. Pedronette, and J. Almeida, "Unsupervised similarity learning through cartesian product of ranking references," *Pattern Recognition Letters*, vol. 114, pp. 41–52, 2018.

[18] R. da S. Torres and A. X. Falcão, "Content-Based Image Retrieval: Theory and Applications," *Revista de Informática Teórica e Aplicada*, vol. 13, no. 2, pp. 161–185, 2006.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database."

[20] O. Kurland, "The cluster hypothesis in information retrieval," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '13, 2013, p. 1126.

[21] D. C. G. Pedronette, O. A. Penatti, and R. d. S. Torres, "Unsupervised manifold learning using reciprocal knn graphs in image re-ranking and rank aggregation tasks," *Image and Vision Computing*, vol. 32, no. 2, pp. 120–130, 2014.

[22] D. C. G. Pedronette and R. d. S. Torres, "Unsupervised effectiveness estimation for image retrieval using reciprocal rank information," in *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, 2015, pp. 321–328.

[23] L. P. Valem and D. C. G. Pedronette, "A denoising convolutional neural network for self-supervised rank effectiveness estimation on image retrieval," in *ICMR '21: Int. Conf. on Multimedia Retrieval*, 2021.

[24] S. Lin, "Rank aggregation methods," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 5, pp. 555–570, 2010.

[25] L. P. Valem, D. C. G. Pedronette, and L. J. Latecki, "Graph convolutional networks based on manifold learning for semi-supervised image classification," *Computer Vision and Image Understanding*, vol. 227, p. 103618, 2023.

[26] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1447–1454.

[27] G.-H. Liu and J.-Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern recognition*, vol. 46, no. 1, pp. 188–198, 2013.

[28] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-ucsd birds 200," 2010.

[29] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," *Advances in neural information proc. systems*, vol. 30, 2017.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[32] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[33] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019.

[34] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[35] L. P. Valem and D. C. G. a. Pedronette, "An unsupervised distance learning framework for multimedia retrieval," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '17, 2017, pp. 107–111.