Unsupervised Effectiveness Estimation Measure Based on Rank Correlation for Image Retrieval

Thiago César Castilho Almeida[0000−0002−2167−0463], Lucas Pascotti Valem[0000−0002−3833−9072], and Daniel Carlos Guimarães Pedronette[0000−0002−2867−4838]

São Paulo State University (UNESP), Rio Claro, SP, Brazil
{tc.almeida,lucas.valem,daniel.pedronette}@unesp.br

Abstract. In recent years, the amount of image data has increased exponentially, driven by advancements in digital technologies. As the volume of data expands, the efforts required for labeling also escalate, which is costly and time-consuming. This scenario highlights the critical need for methods capable of delivering effective results in scenarios with few or no labels at all. In unsupervised retrieval, the task of Query Performance Prediction (QPP) is crucial and challenging, as it involves estimating the effectiveness of a query without labeled data. Besides promising, the QPP approaches are still largely unexplored for image retrieval. Additionally, recent approaches require training and do not exploit rank correlation to model the data. To address this gap, we propose a novel QPP measure named Accumulated JaccardMax, which considers contextual similarity information and innovates by exploiting a recent rank correlation measure to assess the effectiveness of ranked lists. It provides a robust estimation by analyzing the ranked lists in different neighborhood depths and does not require any training or labeled data. Extensive experiments were conducted across 5 datasets and over 20 different features including hand-crafted (e.g., color, shape, texture) and deep learning (e.g., Convolutional Networks and Vision Transformers) models. The results reveal that the proposed unsupervised measure exhibits a high correlation with the Mean Average Precision (MAP) in most cases, achieving results that are better or comparable to the baseline approaches in the literature.

Keywords: Query Performance Prediction · Image Retrieval.

1 Introduction

The widespread adoption of the Internet and digital devices, such as cameras and smartphones, has dramatically increased the volume of multimedia data available online. This shift has transformed the population into both consumers and producers of content. Consequently, the search task has become even more important, being used as a tool to group and organize the data. In this scenario, Content-Based Image Retrieval (CBIR) systems [27] are getting attention.

The CBIR systems can rank and retrieve images according to their visual properties, removing possible ambiguities caused by homonyms in the language.

This is useful because user intervention is needed to surpass this problem when performing text-based image retrieval, while CBIR systems are fully automated. Traditional CBIR systems commonly use shape, color, or texture feature descriptors. However, a decade ago, the emergence of deep learning strategies caused a shift in the retrieval systems, making the Convolutional Neural Networks and Vision Transformers the state-of-the-art strategies for feature extraction [6].

Due to the ease of content creation and sharing, a vast amount of data remains unlabeled, emphasizing the critical need for unsupervised learning. To address this challenge, recent works have been exploiting contextual similarity information [22] within image datasets to perform various tasks in unsupervised retrieval. These tasks include re-ranking approaches [36,40], and rank-aggregation [35,41]. By considering the encoded information within the data, these methods uncover valuable insights without labels.

This work is about Query Performance Prediction (QPP), also referred to in the literature as effectiveness estimation, query difficulty prediction, and query difficulty estimation. A QPP task is responsible for estimating the effectiveness of a ranked list, distinguishing effective queries from poor ones. Initially developed for traditional text-based information retrieval systems, applying QPP to image retrieval tasks is still a largely unexplored area [25].

Therefore, recent works are exploring QPP techniques in the domain of image retrieval [20,26,34,37]. Given the importance of accurately modeling information in these scenarios, some measures like the Authority [23], Reciprocal Density [21], and Full-Intersection [26] measures were proposed to perform in a fully unsupervised way. Modeled in a graph-based structure, these measures are grounded on the cluster hypothesis [12], which considers that images highly ranked in a ranked list should appear in the ranked lists of each other. Alternatively, recent self-supervised approaches were proposed, the Deep Rank Noise Estimator [34] and Regression for Query Performance Prediction Framework [37]. Both strategies use synthetic data for training and incorporate innovative techniques to exploit the contextual information within an image collection.

This work introduces a novel unsupervised effectiveness estimation measure for image retrieval tasks, the Accumulated JaccardMax. The main contributions are as follows: *(i)* It innovates by exploiting a recent rank correlation measure, the JaccardMax [33], which is known for its robustness to neighborhood size and its ability to exploit contextual similarity information at various depths of ranked lists; *(ii)* It is fully unsupervised, eliminating the need for training, and is based on the hypothesis that images with similar ranked lists tend to be similar too [12]; *(iii)* It is employed for image retrieval, a largely unexplored area for most approaches in this category [25]; *(iv)* Experimental results conducted on five public datasets have shown that the proposed measure achieved great results compared to the baselines, outperforming even robust self-supervised strategies.

The remainder of this paper is organized as follows: Section 2 details the problem formulation, presenting the formal definitions and the main concepts. Section 3 presents the proposed measure. Section 4 reports and discusses the

experimental results (**a link**[1] **is added for the supplementary material and the source code of our approach**). Section 5 concludes this work.

2 Background

This section presents the concepts used in this work. Section 2.1 presents the problem formulation, defining the rank and retrieval models. Section 2.2 describes the unsupervised effectiveness estimation used as baselines.

2.1 Formal Definition

This section conceptually defines the problem. Let $\mathcal{C} = \{img_1, img_2, \ldots, img_n\}$ be an image collection and $n = |\mathcal{C}|$ the collection size. Let $\mathcal{D}$ be an image descriptor of shape, color, texture or a deep learning technique defined by a tuple $\mathcal{D} = (\epsilon, \rho)$, where $\epsilon : img_i \rightarrow \mathbb{R}^d$ is the function that extracts the image's features, and $\rho : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^+$ is the function that computes the distance between the images considering their features.

Let x_i be the spatial representation of the image img_i in the $\mathbb{R}^d$ space, given by the d features extracted from the descriptor $\mathcal{D}$. The distance between two images can be defined as $\rho(x_i, x_j)$. Euclidean-like distance functions are commonly used to compute the image's distance. The distance given by the ρ function is the original distance between two objects in the image collection. The notation $\rho(i, j)$ is used along the paper for readability.

The image rankings are represented by ranked lists. A ranked list $\tau_q = (img_1, img_2, \ldots, img_n)$ is a permutation of the image collection. The rank of an image img_i in the ranked list τ_q is defined as $\tau_q(i)$. If $\tau_q(i) < \tau_q(j)$, then $\rho(q, i) \leq \rho(q, j)$, meaning that the distance between img_q and img_i is less than the distance between img_q and img_j. This guarantees the most similar query images are on the top positions of the ranked lists. Every image in the collection has its own ranked list. The set of all ranked lists can be defined as $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_n\}$.

The neighborhood set that contains the most k similar images to a query img_q can be defined as $\mathcal{N}(q, k)$, where:

$$\mathcal{N}(q,k) = \{\mathcal{X} \subseteq \mathcal{C}, |\mathcal{X}| = k \ \wedge \ \forall \ img_i \in \mathcal{X}, img_j \in \mathcal{C} - \mathcal{X} : \tau_q(i) < \tau_q(j)\}. \quad (1)$$

2.2 Effectiveness Estimation Approaches

There are two main categories of QPP approaches [25]: pre-retrieval and post-retrieval. This work focuses on post-retrieval QPP. In contrast to pre-retrieval, post-retrieval methods predict the quality of the query after the retrieval process has taken place. In other words, these approaches are responsible for estimating the effectiveness of a ranked list in a scenario with unlabeled data. This section formally defines the four baselines used in this work: two are unsupervised measures, while the other two are recent self-supervised methods.

[1] **Supplementary files and source code:** accjacmax.lucasvalem.com

1) Authority Score: The Authority Score [23] is a graph-based unsupervised effectiveness estimation measure. Given a query image img_q, each image in $\mathcal{N}(q,k)$ is a node. For each $img_j \in \mathcal{N}(q,k) \ \wedge \ img_j \in \mathcal{N}(i,k)$, where $img_i \in \mathcal{N}(q,k)$, an edge is created between img_q and img_i. The Authority Score counts how many edges were created. In other words, it computes the graph density. The Authority Score for a query image img_q can be defined as:

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

where $f_{in}(j,q)$ returns 1 if img_j is between the k-nearest neighborhoods of image img_q and 0 otherwise.

2) Reciprocal Neighborhood Density: The Reciprocal Density score [21] is based on the number of reciprocal neighbors between two rankings. Formally, if $img_i \in \mathcal{N}(q,k) \wedge img_q \in \mathcal{N}(i,k)$, the score is incremented, which is given by:

$$n_r(q,i) = \frac{\sum_{j \in \mathcal{N}(q,k)} \sum_{l \in \mathcal{N}(i,k)} f_r(j,l) \times w_r(q,j) \times w_r(i,l)}{k^4}, \tag{3}$$

where $f_r(j,l) \rightarrow \{0,1\}$ returns 1 if $img_j \in \mathcal{N}(q,k)$ and 0 otherwise; and $w_r(q,j) = k+1-\tau_q(j)$ is the weight function. The higher the weight, the more occurrences of reciprocal neighborhoods in the top positions of the ranked lists.

3) Deep Rank Noise Estimator (DRNE): The DRNE [34] models ranked lists as grayscale images. Each ranked list contains a level of noise, associated with its effectiveness. Following this idea, the more effective a ranked list, the less noise is associated with it. The method predicts a score for each ranked list associated with the noise level according to its grayscale image. The noise level of a ranked list is encoded by a denoising Convolutional Neural Network (CNN), trained with synthetic data to keep the entire workflow unsupervised.

4) Regression for QPP Framework (RQPPF): The RQPPF [37] also relies on the use of synthetic data, generating meta-features to train a regression model from them. These meta-features, which include information on reciprocal neighborhoods, effectiveness estimation measures, and average ranking positions, are then computed for the real data. The real data is subsequently used as the testing dataset for the regression model. In this work, we report the results using Support Vector Regression (SVR), as it is among the best for this approach.

3 Proposed Method

This section describes the proposed QPP measure named Accumulated JaccardMax. While Subsection 3.1 conceptually defines ranking correlation and the measure used in this work, Subsection 3.2 presents the proposed measure.

3.1 Rank Correlation and JaccardMax

Ranking correlation measures are responsible for determining the similarity between two ranked lists. Based on the proposed formal definition, a ranking correlation measure can be defined as a function $r_c : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R}$, that computes a

value between 0 and 1, where 0 means that the ranked lists are totally different, and 1 meaning that the ranked lists are totally equaled. Ranking correlation measures commonly evaluate the similarity between two ranked lists considering the top-k positions, or the set $\mathcal{N}(img_q, k)$ for an image img_q with ranked list τ_q.

The JaccardMax [33] measure considers the highest Jaccard index value obtained until a depth k as the similarity value between two ranked lists. The main idea of this measure is that a high overlap between two ranked lists, at any depth, is high evidence of similarity between them. The JaccardMax measure between two images img_i and img_j, with τ_i and τ_j ranked lists, respectively, can be formally defined as:

$$JacMax(\tau_i, \tau_j, k) = \max_{1 \leq k_d \leq k} \frac{|\mathcal{N}(img_i, k_d) \cap \mathcal{N}(img_j, k_d)|}{|\mathcal{N}(img_i, k_d) \cup \mathcal{N}(img_j, k_d)|}. \quad (4)$$

3.2 Accumulated JaccardMax

The proposed effectiveness estimation measure, called Accumulated JaccardMax (Acc. JacMax), departs from other methods by not only assessing the reciprocal presence of items in their ranked lists but also incorporating a measure of similarity between them. Unlike earlier approaches that focused solely on list reciprocity, the Accumulated JaccardMax measure utilizes the JaccardMax correlation measure to evaluate the ranked list similarity up to a depth of k, assigning higher weights to similarities found in top positions. This method aims to provide a more comprehensive and accurate evaluation of ranked lists by considering both the similarity information and the position of an image in the ranked list of the query image. The innovation lies in the premise that substantial overlap between ranked lists indicates image similarity, particularly when analyzing the top positions. Thus, a high correlation between the query's ranked list and the ranked list of a k-nearest neighborhood implies a highly effective query ranked list, suggesting Accumulated JaccardMax as a promising strategy for effectiveness estimation in scenarios involving unlabeled data.

The Acc. JacMax score for a query img_q with ranked list τ_q can be given by:

$$AccJacMax(q, k, \alpha) = \frac{\sum_{j \in \mathcal{N}(q,k)} JacMax(\tau_q, \tau_j, k) \times \alpha^{\tau_q(j)}}{k}, \quad (5)$$

where k is the neighborhood size and α is a value in the $[0, 1]$ interval that denotes the weight of each image in the ranked list τ_q. The higher the α, the greater the weight assigned to the top positions. Figure 1 illustrates the computation of the Accumulated JaccardMax score for an image img_q.

4 Experimental Evaluation

This section presents the experimental evaluation conducted. Section 4.1 describes the experimental protocol, the image datasets, descriptors, and other details. Finally, Section 4.2 presents the obtained results.

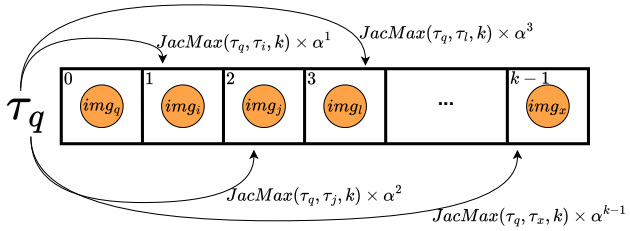


Fig. 1: Computation of Accumulated JaccardMax score for the ranked list τ_q. The images highlighted in orange represent the nearest neighbors of img_q.

4.1 Experimental Protocol

The experiments were conducted in five different datasets described as follows:

- **Soccer [38]:** images from 7 different soccer teams, with 40 images per class, totaling 280 images. Color descriptors are evaluated.
- **Flowers [18]:** distributed by the University of Oxford, this dataset totals 1360 flower images, divided by 17 classes, 80 images per class.
- **MPEG-7 [13]:** containing 1400 images, this shape dataset is divided into 70 classes, 20 images per class. Shape descriptors are evaluated.
- **Brodatz [2]:** composed of 111 different textures, each one is divided into 16 blocks, totaling 1776 images. Texture descriptors are evaluated.
- **Corel5k [16]:** dataset with various images, like animals, tiles, and stained glass. Contains 5000 images, distributed into 50 classes, 100 images per class.

The proposed measure and baseline QPP approaches were compared using the Mean Average Precision (MAP) as the ground truth for evaluating effectiveness. For this evaluation, the MAP was computed using the full ranked list size. To assess the relationship between these measures and MAP, the Pearson correlation coefficient was used, as is commonly done in the literature [25,34,37].

The Pearson correlation coefficient admits values in the $[-1, +1]$ interval, where 1 means a perfect positive linear correlation, and -1 indicates a perfect negative linear correlation. In this scenario, a high positive correlation between the MAP and the effectiveness estimation measure indicates a high similarity between both measures. The higher the correlation between the effectiveness estimation measure and the MAP, the better the measure to evaluate the effectiveness of a ranked list in an unsupervised scenario.

For all datasets, each image was considered as a query. The descriptors used for each dataset were selected considering the properties of them. In total, over 20 different descriptors were used, including hand-crafted (e.g., color, shape, texture) and deep learning (e.g., Convolutional Networks and Vision Transformers) models. All the deep learning feature extractors models were trained on the ImageNet [4] dataset.

Regarding baselines, the Authority [23] and Reciprocal [21] were executed considering an implementation publicly available [35]. For DRNE [34] and RQPPF [37], this work used the results reported in their respective publications due to the

complexity of these methods and their requirement for training on synthetic data. To ensure a fair comparison, the same neighborhood value k was maintained for each dataset across all approaches. Specifically, k was set to 20 for the Soccer, MPEG-7, and Brodatz datasets, reflecting their smaller class sizes. For the Flowers and Corel5k datasets, which feature larger class sizes, k was set to 80. These parameters are also consistent with the ones reported by DRNE [34] and RQPPF [37]. Additionally, we explored a range of α values for the proposed Accum. JaccardMax measure, varying from 1.0, where all JaccardMax values are equally weighted, to 0.90, where higher weights are assigned to the top positions.

4.2 Results

To evaluate our proposed measure against QPP baselines, Table 1 presents the Pearson correlation between MAP and the QPP approaches across all five datasets and various descriptors. Since RQPPF has different variations, the results reported here are the best ones according to its publication [37]. Gray highlights the best value per row, and bold highlights the best value of the proposed measure (Acc. JacMax). The original MAP is reported for reference and the descriptors are sorted based on it, indicating higher correlation values are typically found in descriptors with high effectiveness and vice versa. Notice that the proposed approach achieved the best results in most cases. It also demonstrates that Acc. JacMax is robust to variations in the parameter α, with $\alpha = 0.95$ and $\alpha = 0.90$ likely being close to the optimal choice in most situations.

The Accumulated JaccardMax score consistently outperformed the baselines in most descriptors, demonstrating its robust effectiveness predictions. As previously discussed, the highest correlation values are observed in the highly effective descriptors, particularly VIT-B16 [5], ResNet [7], and ResNeXt [39]. The proposed measure achieved the best average result in 3 out of the 5 datasets. Even for the descriptors where it did not achieve the best results, the difference from the top measure was minimal, ensuring competitiveness in these cases. Also, we highlight that, according to the results of a Student's t-test with 95% confidence, all p-values for the Acc. JaccardMax are below 0.01, providing strong evidence against the null hypothesis in the statistical test.

The results were organized into bar graphs to facilitate comparison between our approach and baseline methods. Figure 2 presents the Pearson correlation between MAP and the QPP approaches for the Brodatz dataset. The red hatched bar represents the Acc. JacMax, showing the results for the best α value reported in Table 1. The results were reported for the Regression for Query Performance Prediction Framework (RQPPF) [37] considering its two variants: "RQPPF + A" and "RQPPF + R", which considered Authority and Reciprocal for computing its meta-features, respectively. The results highlight the superior performance of the proposed measure, underscoring its effectiveness for texture descriptors.

Aiming to visually compare the correlation of the proposed approach with MAP concerning baselines, Figure 3 presents a visual analysis of the unsupervised effectiveness measures and the MAP. Each point in the graphs corresponds to a ranked list. The MAP values of the ranked lists are presented on the x-axis,

Table 1: Pearson correlation (between MAP and QPP, the higher the better) for our proposed approach compared to baselines on different datasets. Gray highlights the best value per row, and bold the best of the proposed measure.

Datasets	Descriptors	Original	Baselines				Acc. JacMax (Ours)			
		MAP	Auth.	Recip.	DRNE	RQPPF	$\alpha = 1$	$\alpha = 0.99$	$\alpha = 0.95$	$\alpha = 0.90$
Flowers	VIT-B16 [5]	87.71%	0.9044	0.9354	—	—	0.9105	0.9229	**0.9409**	0.9334
	CNN-FBResNet [7]	52.56%	0.7374	0.6715	0.7992	—	0.7997	0.8183	**0.8514**	0.8460
	CNN-ResNeXt [39]	51.91%	0.7657	0.6653	0.7927	—	0.8102	0.8285	**0.8646**	0.8613
	CNN-ResNet [7]	51.83%	0.7298	0.6367	0.7990	—	0.7896	0.8086	**0.8420**	0.8331
	CNN-Xception [3]	47.31%	0.7437	0.6406	0.7696	—	0.7741	0.7945	0.8403	**0.8405**
	CNN-SENet [8]	43.16%	0.5872	0.5720	0.6308	—	0.6261	0.6465	**0.6854**	0.6708
	CNN-InceptRN [30]	42.20%	0.6273	0.5336	0.5504	—	0.6670	0.6773	**0.6923**	0.6788
	CNN-BnVGGNet [17]	41.87%	0.4852	0.3618	0.6313	—	0.5916	0.6270	0.6986	**0.7052**
	CNN-NASNetLg [42]	40.74%	0.6309	0.5510	0.5497	—	0.6572	0.6671	**0.6767**	0.6582
	CNN-VGGNet [17]	39.05%	0.5050	0.3284	0.6385	—	0.6131	0.6516	0.7233	**0.7296**
	Average	49.83%	0.6717	0.5896	0.6846	—	0.7239	0.7442	**0.7816**	0.7757
Corel5k	VIT-B16 [5]	75.26%	0.8668	0.8438	—	—	0.8583	**0.8634**	0.8584	0.8375
	CNN-ResNet [7]	64.86%	0.8497	0.8117	—	—	0.8472	**0.8518**	0.8388	0.8023
	CNN-FBResNet [7]	64.25%	0.8546	0.8084	—	—	0.8513	**0.8560**	0.8473	0.8167
	CNN-InceptRN [30]	61.31%	0.8055	0.8106	—	—	0.7917	**0.7925**	0.7630	0.7119
	CNN-ResNeXt [39]	62.45%	0.8702	0.8255	—	—	0.8671	**0.8715**	0.8596	0.8247
	CNN-SENet [8]	57.10%	0.7835	0.8248	—	—	0.7840	**0.7881**	0.7670	0.7203
	CNN-Xception [3]	54.60%	0.8784	0.8493	—	—	0.8707	**0.8737**	0.8546	0.8163
	CNN-NASNetLg [42]	53.78%	0.8125	0.8444	—	—	0.8274	**0.8345**	0.8199	0.7758
	CNN-BnVGGNet [17]	52.82%	0.8206	0.7801	—	—	0.8369	0.8461	**0.8501**	0.8264
	CNN-VGGNet [17]	47.99%	0.8071	0.7709	—	—	0.8289	0.8387	**0.8391**	0.8117
	Average	59.44%	0.8349	0.8170	—	—	0.8364	**0.8416**	0.8298	0.7944
MPEG-7	AIR [9]	89.39%	0.7639	0.7507	0.8771	0.8482	0.8159	0.8243	0.8501	**0.8612**
	ASC [15]	85.28%	0.7659	0.8143	0.7468	0.8293	0.7987	0.8010	**0.8067**	0.8038
	IDSC [14]	81.70%	0.7783	0.8091	0.7477	0.8243	0.7993	0.8013	**0.8063**	0.8033
	CFD [24]	80.71%	0.7982	0.8362	0.8259	0.8589	0.8146	0.8174	**0.8251**	0.8245
	BAS [1]	71.52%	0.7903	0.8401	0.7970	0.8493	0.8261	0.8287	0.8360	**0.8366**
	SS [32]	37.67%	0.7847	0.8132	0.8403	0.8419	0.8293	0.8333	0.8451	**0.8497**
	Average	74.38%	0.7804	0.8106	0.8058	0.8420	0.8140	0.8177	0.8282	**0.8298**
Brodatz	LAS [31]	75.15%	0.6473	0.6348	0.6958	0.7056	0.7819	0.7882	0.8092	**0.8258**
	CCOM [11]	57.57%	0.6354	0.6043	0.6563	0.6820	0.7419	0.7481	0.7679	**0.7815**
	LBP [19]	48.40%	0.4961	0.4221	0.4998	0.5214	0.6206	0.6273	0.6491	**0.6648**
	Average	60.37%	0.5929	0.5537	0.6173	0.6363	0.7148	0.7212	0.7421	**0.7574**
Soccer	BIC [28]	39.38%	0.4507	0.3650	—	—	0.4532	0.4562	**0.4637**	0.4615
	ACC [10]	37.28%	0.4919	0.4061	—	—	0.4928	0.4950	**0.4992**	0.4938
	GCH [29]	32.21%	0.2259	0.1664	—	—	**0.2131**	0.2110	0.2000	0.1820
	Average	36.29%	0.3895	0.3125	—	—	0.3864	0.3874	**0.3876**	0.3791

and the effectiveness estimation values give the values on the y-axis. The higher the correlation, the more linear the behavior tends to be. From the, we can see that the Acc. JacMax presents a more linear behavior than Reciprocal.

Finally, Figure 4 presents ranked lists to illustrate the prediction of the Accumulated JaccardMax measure compared to the MAP. The examples are shown for the Corel5k dataset. Query images are highlighted in green, while images from different classes than the query are highlighted in red. A good query and a bad query were selected. Notice that the proposed measure exhibits a higher score for lists with a higher MAP, and vice versa.

5 Conclusion

This work introduced a novel unsupervised effectiveness estimation measure named Accumulated JaccardMax for image retrieval tasks based on ranking

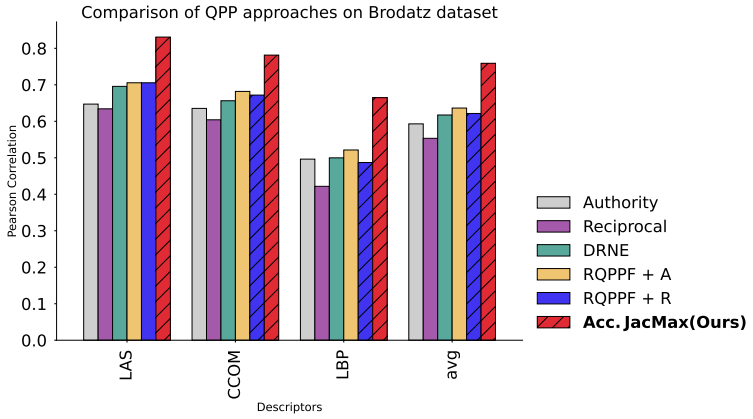
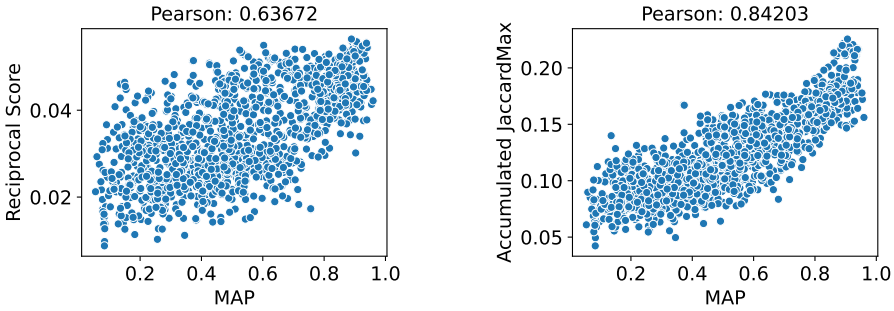


Fig. 2: Pearson correlation between MAP and various QPP approaches, including baselines and our proposed measure (Acc. JacMax) on the Brodatz dataset.

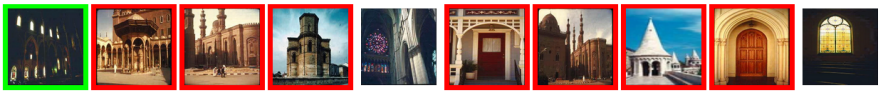


(a) Reciprocal score vs. MAP.

(b) Acc. JacMax ($\alpha = .95$) vs. MAP.

Fig. 3: Correlation of MAP and QPP on Flowers dataset with ResNet descriptor.

Ranked list with $MAP = 92.91\%$ and $Acc.\ JacMax\ Score = 0.21124$



Ranked list with $MAP = 7.21\%$ and $Acc.\ JacMax\ Score = 0.17414$

Fig. 4: Examples of ranked lists along with their MAP value and Accumulated JaccardMax score on the Corel5k dataset with ResNeXt descriptor.

correlation. The results revealed that the proposed measure outperformed QPP approaches in most cases across various descriptors and datasets. Even in cases where it did not surpass the baseline measures, it remained competitive. For evaluation, the MAP measure was used as the ground truth for both the proposed and baseline measures, serving to assess the correlation between the supervised and unsupervised measures.

Future work will focus on evaluating the performance of Accumulated JaccardMax in combination with other baseline measures to enhance performance, assessing its effectiveness in rank fusion and aggregation tasks, and exploring its impact on additional image retrieval tasks. We also intend to evaluate it in other domains (e.g., multimedia retrieval such as video and sound).

Acknowledgments. The authors are grateful to the São Paulo Research Foundation - FAPESP (grant #2018/15597-6), the Brazilian National Council for Scientific and Technological Development - CNPq (grants #313193/2023-1 and #422667/2021-8), and Petrobras (grant #2023/00095-3) for their financial support.

References

1. Arica, N., Yarman Vural, F.T.: BAS: a perceptual shape descriptor based on the beam angle statistics. Pattern Recognition Letters **24**(9-10), 1627–1639 (Jun 2003)
2. Brodatz, P.: Textures: A Photographic Album for Artists and Designers. Dover books on art, graphic art, handicrafts, Dover Publications (1966)
3. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1800–1807 (2017)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv **abs/2010.11929** (2020)
6. Dubey, S.R.: A decade survey of content based image retrieval using deep learning. IEEE Trans. on Circuits and Systems for Video Tech. **32**(5), 2687–2704 (2022)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
8. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Trans. on Pattern Analysis and Machine Intelligence **42**(8), 2011–2023 (2020)
9. Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J.M., Mattern, F., Mitchell, J.C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M.Y., Weikum, G., Gopalan, R., Turaga, P., Chellappa, R.: Articulation-Invariant Representation of Non-planar Shapes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) Computer Vision – ECCV 2010, vol. 6313, pp. 286–299. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
10. Jing Huang, Kumar, S., Mitra, M., Wei-Jing Zhu, Zabih, R.: Image indexing using color correlograms. In: Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 762–768. IEEE, San Juan, Puerto Rico (1997)

11. Kovalev, V., Volmer, S.: Color co-occurrence descriptors for querying-by-example. In: Proceedings 1998 MultiMedia Modeling. MMM'98 (Cat. No.98EX200). pp. 32–38 (Oct 1998)
12. Kurland, O.: The cluster hypothesis in information retrieval. In: de Rijke, M., Kenter, T., de Vries, A.P., Zhai, C., de Jong, F., Radinsky, K., Hofmann, K. (eds.) Advances in Information Retrieval. pp. 823–826. Springer International Publishing, Cham (2014)
13. Latecki, L., Lakamper, R., Eckhardt, T.: Shape descriptors for non-rigid shapes with a single closed contour. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662). vol. 1, pp. 424–429 vol.1 (Jun 2000), iSSN: 1063-6919
14. Ling, H., Jacobs, D.W.: Shape Classification Using the Inner-Distance. IEEE Trans. on Pattern Analysis and Machine Intelligence **29**(2), 286–299 (Feb 2007)
15. Ling, H., Yang, X., Latecki, L.J.: Balancing deformability and discriminability for shape matching. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) Computer Vision – ECCV 2010. pp. 411–424. Springer, Berlin, Heidelberg (2010)
16. Liu, G.H., Yang, J.Y.: Content-based image retrieval using color difference histogram. Pattern Recognition **46**(1), 188–198 (Jan 2013)
17. Liu, S., Deng, W.: Very deep convolutional neural network based image classification using small training sample size. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). pp. 730–734 (2015)
18. Nilsback, M.E., Zisserman, A.: A Visual Vocabulary for Flower Classification. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1447–1454. IEEE, New York, NY, USA (2006)
19. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7), 971–987 (Jul 2002)
20. Oliveira, A., Rocha, A.: Contextual features and sequence labeling techniques for relevance prediction in retrieval. In: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP). pp. 305–310 (2020)
21. Pedronette, D.C.G., Penatti, O.A.B., Calumby, R.T., da Silva Torres, R.: Unsupervised distance learning by reciprocal knn distance for image retrieval. In: Kankanhalli, M.S., Rüger, S.M., Manmatha, R., Jose, J.M., van Rijsbergen, K. (eds.) International Conference on Multimedia Retrieval, ICMR '14, Glasgow, United Kingdom - April 01 - 04, 2014. p. 345. ACM (2014)
22. Pedronette, D.C.G., Torres, R.d.S.: Exploiting contextual information for image re-ranking and rank aggregation. International Journal of Multimedia Information Retrieval **1**, 115–128 (2012)
23. Pedronette, D.C.G., Penatti, O.A., da S. Torres, R.: Unsupervised manifold learning using reciprocal knn graphs in image re-ranking and rank aggregation tasks. Image and Vision Computing **32**(2), 120–130 (2014)
24. Pedronette, D.C.G., Torres, R.d.S.: Shape Retrieval using Contour Features and Distance Optimization. In: Proceedings of the International Conference on Computer Vision Theory and Applications. pp. 197–202. SciTePress - Science and and Technology Publications, Angers, France (2010)
25. Poesina, E., Ionescu, R.T., Mothe, J.: iqpp: A benchmark for image query performance prediction. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 2953–2963. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023)

26. Presotto, J.G.C., Valem, L.P., Pedronette, D.C.G.: Unsupervised effectiveness estimation through intersection of ranking references. In: Computer Analysis of Images and Patterns - CAIP 2019. vol. 11679, pp. 231–244 (2019)
27. Shriram K. Vasudevan, P.L.K. Priyadarsini, S.V.: Content Based Image Retrieval (CBIR): A deeper insight. LAP LAMBERT Academic Publishing (2012)
28. Stehling, R.O., Nascimento, M.A., Falcão, A.X.: A compact and efficient image retrieval approach based on border/interior pixel classification. In: Proceedings of the eleventh international conference on Information and knowledge management. pp. 102–109. ACM, McLean Virginia USA (Nov 2002)
29. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision **7**(1), 11–32 (Nov 1991)
30. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proc. of the 31st AAAI Conference on Artificial Intelligence. p. 4278–4284. AAAI'17, AAAI Press (2017)
31. Tao, B., Dickinson, B.W.: Texture Recognition and Image Retrieval Using Gradient Indexing. Journal of Visual Communication and Image Representation **11**(3), 327–342 (Sep 2000)
32. Torres, R.D.S., Falcão, A.: Contour salience descriptors for effective image retrieval and analysis. Image and Vision Computing **25**(1), 3–13 (Jan 2007)
33. Valem, L.P., Kawai, V.A.S., Pereira-Ferrero, V.H., Pedronette, D.C.G.: A Novel Rank Correlation Measure for Manifold Learning on Image Retrieval and Person Re-ID. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 1371–1375. IEEE, Bordeaux, France (Oct 2022)
34. Valem, L.P., Pedronette, D.C.G.a.: A denoising convolutional neural network for self-supervised rank effectiveness estimation on image retrieval. In: Proceedings of the 2021 International Conference on Multimedia Retrieval. p. 294–302. ICMR '21, Association for Computing Machinery, New York, NY, USA (2021)
35. Valem, L.P., Pedronette, D.C.G.: Unsupervised selective rank fusion for image retrieval tasks. Neurocomputing **377**, 182–199 (2020)
36. 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 **32**, 2811–2826 (2023)
37. Valem, L.P., Pereira-Ferrero, V.H., Pedronette, D.C.G.: Self-supervised regression for query performance prediction on image retrieval. In: 2023 IEEE Sixth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). pp. 95–98. IEEE Computer Society, Los Alamitos, CA, USA (sep 2023)
38. Van De Weijer, J., Schmid, C.: Coloring Local Feature Extraction. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) Computer Vision – ECCV 2006, vol. 3952, pp. 334–348. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
39. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5987–5995 (2017)
40. Zhang, Y., Qian, Q., Wang, H., Liu, C., Chen, W., Wang, F.: Graph convolution based efficient re-ranking for visual retrieval. IEEE Transactions on Multimedia **26**, 1089–1101 (2024)
41. Zheng, L., Wang, S., Tian, L., Fei He, Liu, Z., Tian, Q.: Query-adaptive late fusion for image search and person re-identification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1741–1750 (2015)
42. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8697–8710 (2018)