Unsupervised Effectiveness Estimation for Image Retrieval using Reciprocal Rank Information

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

Ricardo da S. Torres
Recod Lab - Institute of Computing
University of Campinas (UNICAMP)
Campinas - Brazil
rtorres@ic.unicamp.br

***Abstract*—In this paper, we present an unsupervised approach for estimating the effectiveness of image retrieval results obtained for a given query. The proposed approach does not require any training procedure and the computational efforts needed are very low, since only the top-k results are analyzed. In addition, we also discuss the use of the unsupervised measures in two novel rank aggregation methods, which assign weights to ranked lists according to their effectiveness estimation. An experimental evaluation was conducted considering different datasets and various image descriptors. Experimental results demonstrate the capacity of the proposed measures in correctly estimating the effectiveness of different queries in an unsupervised manner. The linear correlation between the proposed and widely used effectiveness evaluation measures achieves scores up to 0.86 for some descriptors.**

***Keywords*-content-based image retrieval; unsupervised effectiveness estimation; query difficult prediction**

I. Introduction

The consistent development of digital image acquisition and sharing technologies led to a huge growth of image collections in last decades. In this scenario, indexing and searching for collection images is of paramount importance. One common approach to support these tasks consider the use of image visual content and the construction of the well-known Content-Based Image Retrieval (CBIR) systems. The main objective of these systems is to retrieve the most similar images ranked according to their similarity to a query input (e.g., query image).

Although the continued research and development in last two decades [1], the CBIR technology still faces several challenges. A common problem faced by all current approaches is their reliance on visual similarity for judging semantic similarity, which may be difficult due to the semantic gap between low-level content features and higher-level concepts [1].

Such inherent difficulties can directly affect the effectiveness of image retrieval systems based on the visual content. The effectiveness of image retrieval tasks is commonly associated with the relevance of top-ranked images regarding to the query image. In addition, various post-processing [2]–[4] methods use the information encoded in top positions of ranked lists for improving the effectiveness results. However, distinct queries present different search difficulty levels, depending on considered visual features. For some queries, the search systems may return effective results, while for others, the search results may be very unsatisfactory [5].

In fact, it would be very desirable estimating the effectiveness of retrieved results without the need of user intervention, e.g., in an unsupervised way. We claim that the quality of retrieved results for a given query may be used to improve the search system automatically. For example, considering that the effectiveness of results for a given query is known, the CBIR system may perform re-ranking [4], [6], [7] or may support relevance feedback [8]–[10] sessions for improving the quality of low-effective queries. On the other hand, a greater relevance can be assigned to high-effective queries, and that information may be used to tune searching models aiming at improving the results associated with future queries.

In textual information retrieval systems, Query Difficulty Prediction (QDP) approaches have been investigated for years [11]–[13], as an attempt to predict the quality of the search result for a query over a given collection [5]. Considering the textual domain, studies have been conducted [13] to investigate the reason why some queries are more difficult than others and models have been proposed for predicting the query difficulty. Ranking robustness [11], which refers to a property of a ranked list of documents that indicates how stable the ranking is in the presence of uncertainty, consistently correlates with query performance in textual domains.

In image retrieval applications, however, the use of unsupervised effectiveness estimation approaches is still little exploited. Tian et al. [5] proposed a query difficulty prediction approach for web image search tasks, evaluated on textual queries. A model to predict the query difficulty through a machine learning approach is employed. Machine learning models are also used for query difficult prediction in contextual image retrieval system [14]. The objective is to annotate the word/phrase in a document with images. Unsupervised approaches were proposed [15] for effectiveness estimation of image-based queries using visual representations of the query neighborhood.

In this paper, we present a completely unsupervised approach for estimating the effectiveness of image retrieval tasks. A reciprocal reference analysis is employed, based on two measures recently proposed for unsupervised distance learning tasks [3], [16]. The approach does not require any training procedure and the computational efforts needed are very low, since only the top-k results are analyzed. In addition, we also present an application of discussed measures, proposing two novel rank aggregation methods, which assign weights to ranked lists according to their effectiveness estimation. An experimental evaluation was conducted considering different datasets and various image descriptors, based on shape, color, and texture features. Experimental results demonstrate the capacity of the proposed measures in correctly estimating the effectiveness of different queries. We show that the linear correlation between the proposed and widely used effectiveness measures achieves scores up to 0.86 for some descriptors.

The remainder of this paper is organized as follows. Section II discusses the image retrieval model. Section III presents the unsupervised effectiveness estimation measures, while Section IV discusses their use in rank aggregation tasks. Section V presents the conducted experimental evaluation, and, finally, Section VI discusses the conclusions and draws future work.

II. Image Retrieval Model

This section briefly defines the image retrieval model adopted along the paper. Let $\mathcal{C}$=$\{img_1, img_2, \ldots, img_n\}$ be an image collection, where $n = |\mathcal{C}|$ defines the size of the collection. Let D be an image descriptor. An image descriptor can be defined [17] as a tuple (ϵ, ρ):

- ϵ: $\hat{I} \rightarrow \mathbb{R}^n$ is a function, which extracts a feature vector $v_{\hat{I}}$ from an image $\hat{I}$;
- ρ: $\mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is a distance function that computes the distance between two images according to the distance between their corresponding feature vectors.

The distance between two images img_i and img_j is given by the value of $\rho(\epsilon(img_i),\ \epsilon(img_j))$. The notation $\rho(i,j)$ is used for readability purposes. A distance matrix A can be computed based on distance among all images, such that $A_{ij} = \rho(i,j)$.

The ranking model adopted is defined based on ranked lists [7]. A ranked list τ_q can be also computed for query image img_q, based on distance ρ. The ranked list τ_q=$(img_1, img_2, \ldots, img_n)$ can be defined as a permutation of the collection $\mathcal{C}$. A permutation τ_q is a bijection from the set $\mathcal{C}$ onto the set $[N] = \{1, 2, \ldots, n\}$. The position (or rank) of image img_i in the ranked list τ_q, is denoted by $\tau_q(i)$. If img_i is ranked before img_j in the ranked list of img_q, that is, $\tau_q(i) < \tau_q(j)$, then $\rho(q,i) \leq \rho(q,j)$.

Taking every image $img_i \in \mathcal{C}$ as a query image img_q, a set of ranked lists $\mathcal{R}$ can be computed as follows:

$$\mathcal{R} = \{\tau_1, \tau_2, \ldots, \tau_n\}. \tag{1}$$

We can also define a neighborhood set that contains the most similar images to img_q as $\mathcal{N}(q,k)$. For the k-nearest neighbor query, we have $|\mathcal{N}(q,k)| = k$, which is formally defined as follows:

$$\mathcal{N}(q,k) = \{\mathcal{S} \subseteq \mathcal{C}, |\mathcal{S}| = k \wedge \forall img_i \in \mathcal{S}, img_j \in \mathcal{C} - \mathcal{S} : \tau_q(i) < \tau_q(j)\}. \tag{2}$$

Based on the discussed image model, next section describes the use of unsupervised measures for effectiveness estimation.

III. Unsupervised Effectiveness Estimation Measures

The ranked lists are a rich source of information upon the whole image collection, since they encode comparison relationships among all images. Our objective is to exploit the information available in different ranked lists for accurately estimating the effectiveness of retrieved results.

The cluster hypothesis [18] states that *"closely associated items tend to be relevant to the same requests."* Therefore, it is expected that images at top positions of a high-effective ranked list refer to each other at the top positions of their own ranked lists [3]. In next sub-sections, we discuss two approaches, based on the analysis of reciprocal references among top positions of ranked lists. The two discussed measures were recently proposed as part of unsupervised distance learning procedures [3], [16]. In this work, we aim at analyzing and evaluating how effective the density of reciprocal references computed by these measures is for predicting the effectiveness of queries.

A. Reciprocal Neighborhood Density

The Reciprocal kNN distance [16] between two images img_q, $img_i \in \mathcal{C}$ is computed based on the number of reciprocal neighbors at top positions of their ranked lists $\tau_q, \tau_i \in \mathcal{R}$. Two images are considered reciprocal neighbors if they are in the neighborhood set of each other (formally, if $img_i \in \mathcal{N}(q,k) \wedge img_q \in \mathcal{N}(i,k)$).

The score based on the number of reciprocal neighbors and its respectively weights are given by the function $n_r(q,i)$, defined as follows:

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

while the function $f_r(j,l) \rightarrow \{0,1\}$ determines if the images img_j, img_l are reciprocal neighbors. A weight is computed according to the function $w_r(q,j) = k + 1 - \tau_q(j)$. The higher the weight, the more frequent is the occurrence of reciprocal neighbors at top positions of ranked lists.

The score based on the number of reciprocal neighbors $n_r(q,i)$ consider two different ranked lists τ_q and τ_i. However, for effectiveness estimation, we are interested in computing the number of reciprocal neighbors for a single ranked list, e.g., the ranked list $\tau(q)$, computed for the query img_q, for which we want to estimate the effectiveness. Therefore, we define a Reciprocal Neighborhood Density score considering the same image for the two inputs:

$$R_s(i) = n_r(i,i). \tag{4}$$

The rationale for using this function relies on the fact that, if the top-k images are similar to each other, they are also expected to be reciprocal neighbors.

B. Authority Measure

The references among images defined by ranked lists can be formally represented by a graph. Let img_q be the query whose effectiveness we want to estimate and let τ_q be the respectively ranked list. Each image in top-k positions of the ranked list τ_q defines a node. For each image img_i in top-k of τ_q, the ranked list τ_i is also analyzed. If there are images in common in ranked lists τ_q and τ_i, an edge is created. The authority score is computed based on the number of created edges. Figure 1 illustrates the computation of the Authority Score [3].

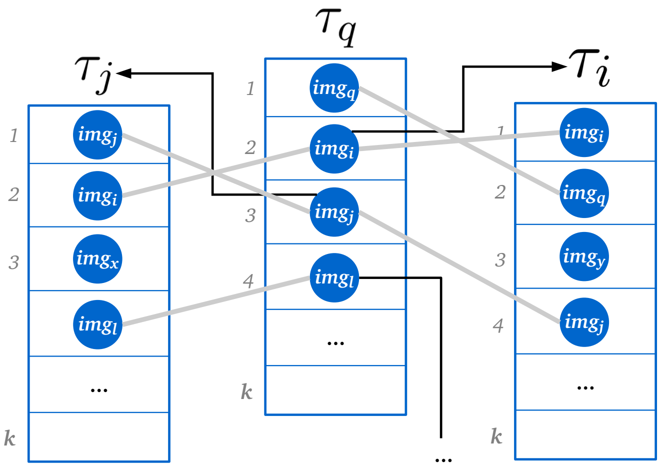

Fig. 1. Example of Authority Score [3] computation for the ranked list τ_q obtained for image img_q. The images img_i, img_j represent k-nearest neighborhs. The Authority Score is defined by the number of references among the k-nearest neighborhs, represented by the edges, in gray.

In other words, the Authority Score [3] measures the density of the graph that represents the references among images at top-k positions. The authority score $A_s(q,k)$ of the ranked list τ_q is formally defined as follows:

$$A_s(q,k) = \frac{\sum_{i \in \mathcal{N}(q,k)} \sum_{j \in \mathcal{N}(i,k)} f_{in}(j,q)}{k^2}, \tag{5}$$

where f_{in} returns 1 if $img_j \in \mathcal{N}(q,k)$ and 0 otherwise.

The score A_s is defined in the interval $[0,1]$. For a full connected neighborhood graph (all k images references each other at top-k positions) this score returns a perfect score $A_s(q,k) = 1$.

IV. Rank Aggregation Methods

Different image retrieval systems and descriptors produce different retrieval results. The information provided by the different ranked lists is commonly complementary, and therefore, have been used for improving the effectiveness of search systems. This is the objective of rank aggregation methods, which aim at combining different ranked lists in order to obtain a more effective one.

Rank aggregation approaches are often unsupervised, presenting the advantage of requiring no training data. On the other hand, without any labeled information, such methods are not capable of distinguishing between high-effective and low-effective ranked lists [15]. In this scenario, the use of unsupervised effectiveness estimation measures can be very useful, as it allows for assigning a relevance score to ranked lists computed by different image descriptors and improving the effectiveness of combined results.

The use of unsupervised effectiveness estimation measures for rank aggregation tasks is illustrated in Figure 2. For each image descriptor, the set of ranked lists is computed, based on feature extraction and distance computation steps. The retrieval results in green and red indicates relevant and non-relevant images. The effectiveness estimation measures are also computed independently of each descriptor. Finally, the rank aggregation method combines the information of both ranked lists and effectiveness estimation measures for producing a final rank.

We propose extensions for two rank aggregation methods: the traditional Borda [19] and the recently proposed Reciprocal Rank Fusion [20] methods. Both Borda [19] and Reciprocal [20] methods consider the rank information, i.e., the positions of images in ranked lists produced by different descriptors. Let $\mathcal{D}$=$\{D_1, D_2, \ldots, D_m\}$ be a set of CBIR descriptors. Let img_q be a query image. Each descriptor $D_j \in \mathcal{D}$ compute a different ranked list τ_{qD_j} for the query image img_q. A given image img_i is ranked at different positions (defined by $\tau_{qD_j}(i)$) according to each descriptor $D_j \in \mathcal{D}$. The rank aggregation methods use these different rank positions aiming at computing a new distance/similarity score between images img_q and img_i.

Different from the original methods and from other initiatives [15], we also exploit the reciprocal rank positions, using the position of the query image img_q in the ranked lists of other images. We formally define the proposed methods in the following.

A. Borda Rank Aggregation Method

The Borda [19] method combines the rank information of each image in different ranked lists computed by different descriptors. Specifically, the distance is scored by the number of images not ranked higher than it in the different ranked lists [21]. The new distance $F_B(q,i)$ between images img_q and img_i is computed as follows:

$$F_B(q,i) = \sum_{j=0}^{m} \tau_{qD_j}(i). \tag{6}$$

The Borda method does not consider the reciprocal position of the query img_q in the ranked list of image img_i (given by $\tau_{qD_j}(i)$). Rank information provided by high-effective and

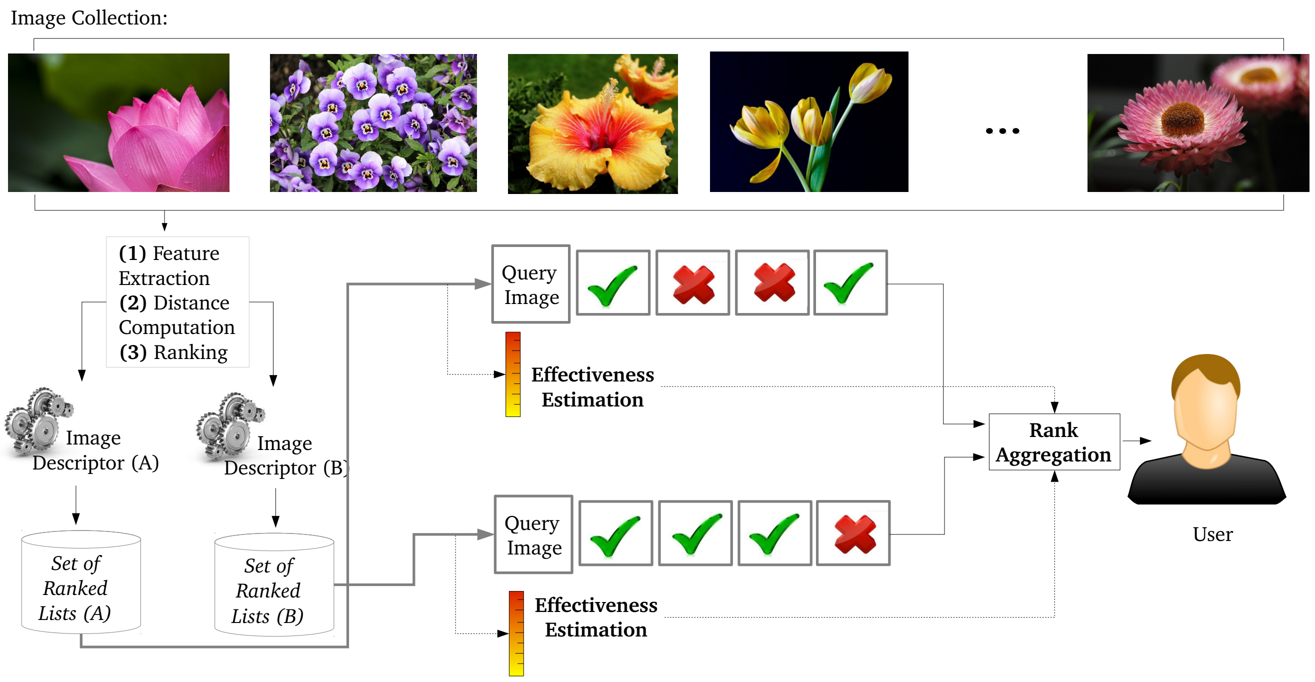

Fig. 2. General workflow of the use of unsupervised estimation measures for rank aggregation tasks.

low-effective ranked lists are also considered in the same way. We propose to use the rank and the reciprocal rank information weighted by the discussed effectiveness estimation measures. The proposed distance is computed as follows:

$$F_{B_W}(q,i) = \sum_{j=0}^{m} (\tau_{qD_j}(i) \times e_{D_j}(q)) + (\tau_{iD_j}(q) \times e_{D_j}(i)), \quad (7)$$

where the function $e_{D_j}(.)$ is an effectiveness estimation measure computed by the Reciprocal Neighborhood Density or the Authority Score.

Notice that the estimation measure is used to determine a weight for position computed for both images img_q and img_i.

B. Reciprocal Rank Fusion

The Reciprocal Rank Fusion uses the rank information for computing a similarity score between images img_q and img_i according to a naive scoring formula:

$$F_R(q,i) = \sum_{j=0}^{m} \frac{1}{k + \tau_{qD_j}(i)}, \quad (8)$$

where k is a constant.

Analogously to the Borda method, the reciprocal rank position (position of img_q in the ranked list of img_i) is not exploited. We also propose to use this information combined with weights defined by the effectiveness estimation measures. The proposed distance is computed as follows:

$$F_{R_W}(q,i) = \sum_{j=0}^{m} \frac{1}{k + (\tau_{qD_j}(i) \times e_{D_j}(q)) + (\tau_{iD_j}(q) \times e_{D_j}(i))}, \quad (9)$$

V. Experimental Evaluation

This section presents conducted experiments for evaluating the accuracy of proposed approaches. We analyzed the measures, considering different descriptors and datasets. We also evaluate the effectiveness of the proposed rank aggregation methods.

A. Datasets and Descriptors

The datasets and image descriptors used in the experimental evaluation are briefly described in the following. Three different datasets and eleven image descriptors, involving shape, color, and texture features are considered.

1) Shape: The MPEG-7 [33] dataset, a well-known shape dataset used for shape descriptors and post-processing methods evaluation, was used in our experiments. The dataset is composed of 1400 shapes divided into 70 classes of 20 images each. Five shape descriptors were considered: Segment Saliences (SS) [22], Beam Angle Statistics (BAS) [23], Inner Distance Shape Context (IDSC) [25], Contour Features Descriptor (CFD) [24], and Aspect Shape Context (ASC) [26].

2) Color: The experiments considering color features were conducted on a dataset used in [34]. The dataset is composed of images from 7 soccer teams, containing 40 images per class. Three color descriptors were considered: Border/Interior Pixel Classification (BIC) [29], Auto Color Correlograms (ACC) [28], and Global Color Histogram (GCH) [27].

3) Texture: The Brodatz [35] dataset, a popular dataset for texture descriptors evaluation was used. The Brodatz dataset is composed of 111 different textures, divided into 16 blocks pixels of non-overlapping sub images, such that 1776 images are considered. Three well-known texture descriptors

TABLE I
PEARSON CORRELATION OF THE EFFECTIVENESS ESTIMATED MEASURES WITH THE AVERAGE PRECISION (AP) MEASURE FOR DIFFERENT DESCRIPTORS AND DATASETS.

Image Descriptor	Type	Dataset	MAP Score	Reciprocal Density	Authority Score	Baselines [15]:	
						NDM	NVDM
SS [22]	Shape	MPEG-7	37.67%	**0.86**	0.82	0.71	0.81
BAS [23]	Shape	MPEG-7	71.52%	**0.86**	0.82	0.79	0.76
CFD [24]	Shape	MPEG-7	80.71%	**0.86**	0.84	0.79	0.67
IDSC [25]	Shape	MPEG-7	81.70%	**0.83**	0.80	0.76	0.60
ASC [26]	Shape	MPEG-7	85.28%	**0.83**	0.77	0.75	0.58
GCH [27]	Color	Soccer	32.24%	0.18	0.22	0.13	**0.26**
ACC [28]	Color	Soccer	37.23%	0.46	**0.52**	0.28	0.49
BIC [29]	Color	Soccer	39.26%	0.41	**0.47**	0.23	0.44
LBP [30]	Texture	Brodatz	48.40%	0.57	**0.63**	0.45	0.54
CCOM [31]	Texture	Brodatz	57.57%	**0.72**	**0.72**	0.08	0.71
LAS [32]	Texture	Brodatz	75.15%	**0.78**	0.77	0.67	0.71
Average				**0.67**	**0.67**	0.51	0.60

were considered in the experiments: Local Binary Patterns (LBP) [30], Color Co-Occurrence Matrix (CCOM) [31], and Local Activity Spectrum (LAS) [32].

B. Effectiveness Estimation Accuracy

This section aims at assessing the accuracy of discussed unsupervised measures for estimating the effectiveness of ranked lists. We evaluated the correlation between the proposed measures and ground-truth measures, as the average precision. We also compared the measures presented in this paper with other two recently proposed unsupervised measures [15], considered as baselines.

1) Experimental Protocol: The Average Precision (AP), which is an effectiveness measure commonly used in information retrieval, is considered as ground-truth measure. Let q be a query item and let N_r be the number of relevant items in a collection for a given query q. Let $\langle r_i | i = 1, 2, \dots, d \rangle$ be a ranked relevance vector to depth d, where r_i indicates the relevance of the ith ranked document scored as either 0 (not relevant) or 1 (relevant), the AP is defined as follows:

$$AP = \frac{1}{N_r} \sum_{i=1}^{d} \left(\frac{r_i}{i} \sum_{j=1}^{i} r_j \right). \tag{10}$$

A statistical measure is used to evaluate the magnitude of a relationship among the effectiveness estimation measures and the average precision. This relationship was evaluated using the Pearson's Correlation Coefficient, defined by:

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}}. \tag{11}$$

Pearson's correlation coefficient r for continuous data ranges from -1 to +1, where $r = 1$ indicates a perfect positive linear relationship and $r = -1$ a perfect decreasing linear relationship. The closer the coefficient is to 1, the stronger the correlation between the variables. The employed protocol is similar to the one used in related work [15], [36].

2) Correlation Results: Table I presents the correlation results for Reciprocal Neighborhood Density and Authority Score considering different descriptors and datasets. We also report the Mean Average Precision (MAP) obtained for each descriptor and the correlation coefficients obtained for considered baselines [15]: Neighborhood Distance Measure (NDM) and Neighborhood Distance Variation Measure (NVDM) measures.

We can observe that the Reciprocal Neighborhood Density measure presents very high correlation coefficients for various descriptors, specially on the MPEG-7 [33] dataset (which presents the higher MAP scores). The Pearson's correlation coefficient achieved 0.86 for some descriptors. The Reciprocal Neighborhood Density assigns higher weights for top positions of ranked lists, justifying its better prediction accuracy for more effective descriptors (with higher MAP scores).

The Authority Score also presents high correlation scores for various descriptors and datasets. However, different from the Reciprocal Neighborhood Density, the descriptors which presented the best results are the descriptors with lower MAP scores. Notice that the Authority Score assigns the same weight for all neighbors, benefiting from the prediction for low-effective descriptors.

Both measures present an average correlation coefficient of 0.67 considering all descriptors and datasets. Considering that the Pearson's correlation coefficient is defined in the interval $[-1, +1]$, we can observe that an accurate prediction of the retrieval effectiveness performance is achieved. The evaluated measures also overcome the baselines for all descriptors, except for the GCH descriptor [27].

C. Rank Aggregation Results

A set of experiments was conducted aiming at evaluating the effectiveness of the proposed rank aggregation methods. The two best descriptors of each dataset (shape, color, and texture) were considered in the experiments.

The Borda and the Reciprocal Rank Fusion were evaluated considering both the Reciprocal Density and the Authority

TABLE II
MAP SCORES OBTAINED FOR THE PROPOSED RANK AGGREGATION METHODS.

Image Descriptor	Type	Dataset	Rank Aggregation Method	Effect. Estimate Measure	MAP Score
CFD [24]	Shape	MPEG-7	-	-	80.71%
ASC [26]			-	-	85.28%
CFD [24]+ASC [26]	Shape	MPEG-7	Borda	-	91.12%
			Borda	Reciprocal Density	94.24%
			Borda	Authority Score	93.93%
			Reciprocal	-	93.80%
			Reciprocal	Reciprocal Density	**95.89%**
			Reciprocal	Authority Score	95.85%
ACC [28]	Color	Soccer	-	-	37.23%
BIC [29]			-	-	39.26%
BIC [29]+ACC [28]	Color	Soccer	Borda	-	38.81%
			Borda	Reciprocal Density	42.49%
			Borda	Authority Score	42.23%
			Reciprocal	-	38.88%
			Reciprocal	Reciprocal Density	**42.51%**
			Reciprocal	Authority Score	42.26%
CCOM [31]	Texture	Brodatz	-	-	57.57%
LAS [32]			-	-	75.15%
LAS [32]+CCOM [31]	Texture	Brodatz	Borda	-	73.92%
			Borda	Reciprocal Density	77.01%
			Borda	Authority Score	77.19%
			Reciprocal	-	75.49%
			Reciprocal	Reciprocal Density	77.94%
			Reciprocal	Authority Score	**78.04%**

Score measures. The results of the proposed rank aggregation methods are showed in Table II, considering MAP scores. The results of the image descriptors in isolation and the original rank aggregation methods (without the use of effectiveness estimation measures) are also presented.

For all datasets, the proposed rank aggregation methods presented the best effectiveness results. As we can observe, the effectiveness results are significantly superior to the original methods and the best image descriptors in isolation.

D. Graphical Correlation Analysis

In this section, we present a graphical analysis of the correlation between the effectiveness estimation measures and the Average Precision (AP). Each point in the graph represents a collection image, where the position in the x axis is defined by the effectiveness estimation measure and the position in the y axis is defined by the average precision of the query.

Figures 3 and 4 illustrate the correlation between effectiveness estimation measures and average precision. Figure 3 presents results for the Reciprocal Density, while Figure 4 shows results for the Authority Score. Both examples considered the MPEG-7 [33] dataset and the CFD [24] shape descriptor. Figures 5 and 6 presents analogous results considering the LAS [32] texture descriptor and the Brodatz [35] dataset, while Figures 7 and 8 considered the ACC [28] color descriptor and the Soccer [34] dataset.

Despite the scale variations between the Reciprocal Density and the Authority Score, we can observe that all graphs approximate a linear relationship with a positive slope, consistent with the high correlation coefficients obtained.

VI. CONCLUSIONS

In this work, two unsupervised measures are used for estimating the effectiveness of retrieval results on image retrieval tasks. Both measures are based on the density of reciprocal references on top positions of ranked lists. The relationships among top retrieval results provides a rich source of information and can be exploited for estimating the effectiveness of image retrieval tasks.

Experiments involving shape, color, and texture descriptors demonstrated that the presented approach can provide accurate prediction of the retrieval effectiveness performance. Novel rank aggregation methods were proposed based on the discussed measures, as an application of unsupervised effectiveness estimation approaches. Very effective results were observed for the proposed rank aggregation methods in various experiments.

Future work focuses on the evaluation of proposed measures for effectiveness estimation in other multimedia retrieval tasks, such as relevance feedback or learning-to-rank approaches.

VII. ACKNOWLEDGMENTS

The authors are grateful to São Paulo Research Foundation - FAPESP (grants 2013/08645-0 and 2013/50169-1), CNPq (grants 306580/2012-8 and 484254/2012-0), CAPES, AMD, and Microsoft Research.

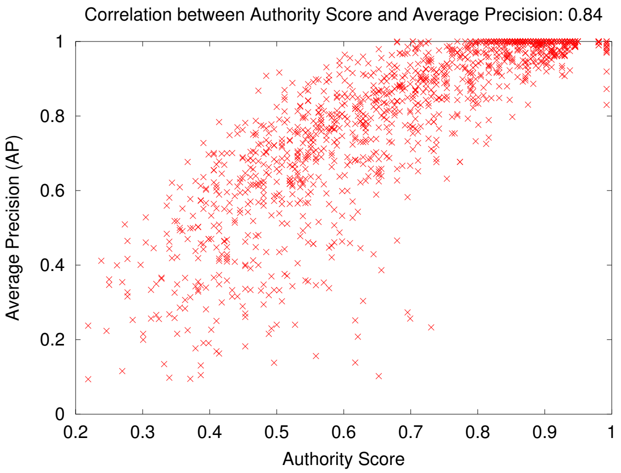

Fig. 3. Correlation Results between Authority Score and Average Precision, considering the CFD [24] shape descriptor.

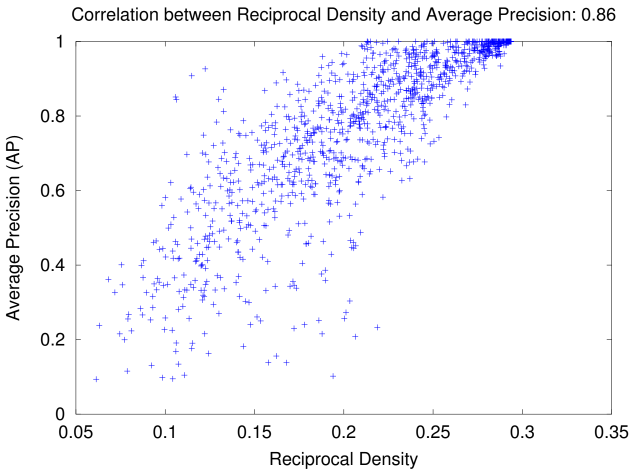

Fig. 4. Correlation Results between Reciprocal Density and Average Precision, considering the CFD [24] shape descriptor.

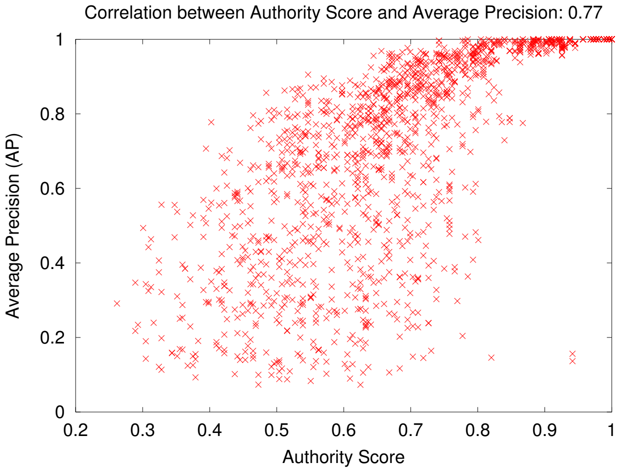

Fig. 5. Correlation Results between Authority Score and Average Precision, considering the LAS [32] texture descriptor.

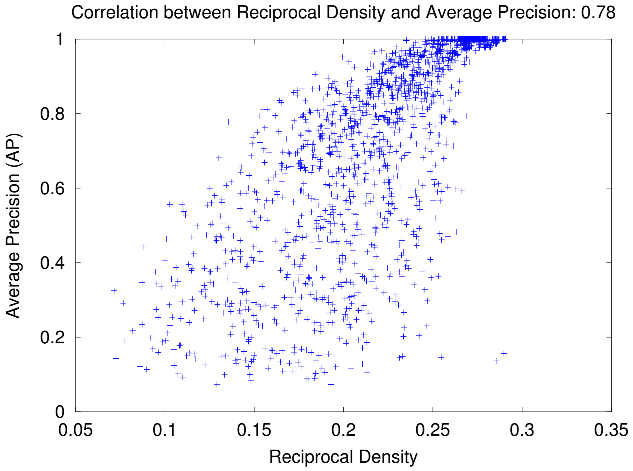

Fig. 6. Correlation Results between Reciprocal Density and Average Precision, considering the LAS [32] texture descriptor.

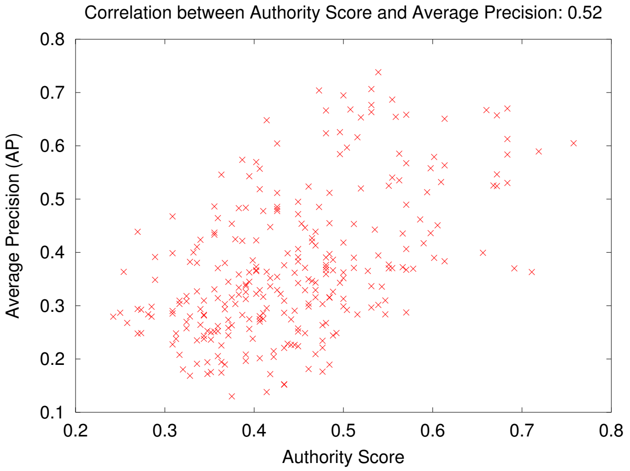

Fig. 7. Correlation Results between Authority Score and Average Precision, considering the ACC [28] color descriptor.

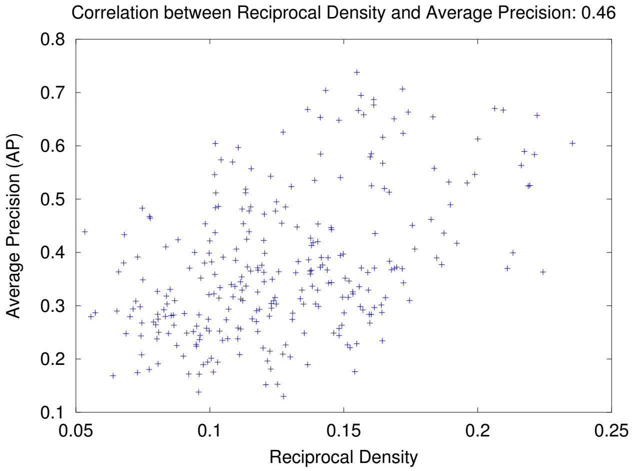

Fig. 8. Correlation Results between Reciprocal Density and Average Precision, considering the ACC [28] color descriptor.

References

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 5:1–5:60, 2008. [Online]. Available: http://doi.acm.org/10.1145/1348246.1348248

[2] X. Yang, S. Koknar-Tezel, and L. J. Latecki, "Locally constrained diffusion process on locally densified distance spaces with applications to shape retrieval." in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2009)*, 2009, pp. 357–364.

[3] D. C. G. Pedronette, O. A. Penatti, and R. da 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.

[4] D. C. G. Pedronette and R. da S. Torres, "Exploiting contextual information for image re-ranking and rank aggregation," *International Journal of Multimedia Information Retrieval*, vol. 1, no. 2, pp. 115–128, 2012.

[5] X. Tian, Y. Lu, and L. Yang, "Query difficulty prediction for web image search," *Multimedia, IEEE Transactions on*, vol. 14, no. 4, pp. 951–962, Aug 2012.

[6] D. C. G. Pedronette and R. da S. Torres, "Using contextual spaces for image re-ranking and rank aggregation," *Multimedia Tools and Applications*, vol. 69, no. 3, pp. 689–716, 2014.

[7] ——, "Image re-ranking and rank aggregation based on similarity of ranked lists," *Pattern Recognition*, vol. 46, no. 8, pp. 2350–2360, 2013.

[8] C. D. Ferreira, J. A. dos Santos, R. da S. Torres, M. A. Gonçalves, R. C. Rezende, and W. Fan, "Relevance feedback based on genetic programming for image retrieval," *Pattern Recogninion Letters*, vol. 32, no. 1, pp. 27–37, 2011.

[9] D. C. G. Pedronette, R. T. Calumby, and R. da S. Torres, "Semi-supervised learning for relevance feedback on image retrieval tasks," in *27th SIBGRAPI Conference on Graphics, Patterns and Images*, Aug 2014, pp. 243–250.

[10] R. T. Calumby, R. da S. Torres, and M. A. Gonçalves, "Multimodal retrieval with relevance feedback based on genetic programming," *Multimedia Tools and Applications*, vol. 69, no. 3, pp. 991–1019, 2014.

[11] Y. Zhou and W. B. Croft, "Ranking robustness: A novel framework to predict query performance," in *ACM Int. Conference on Information and Knowledge Management (CIKM'06)*, 2006, pp. 567–574.

[12] ——, "Query performance prediction in web search environments," in *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*, 2007, pp. 543–550.

[13] D. Carmel, E. Yom-Tov, A. Darlow, and D. Pelleg, "What makes a query difficult?" in *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*, 2006, pp. 390–397.

[14] X. Xing, Y. Zhang, and M. Han, "Query difficulty prediction for contextual image retrieval," in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science, 2010, vol. 5993, pp. 581–585.

[15] D. C. G. Pedronette and R. da S Torres, "Unsupervised measures for estimating the effectiveness of image retrieval systems," in *26th SIBGRAPI - Conference on Graphics, Patterns and Images*, 2013, pp. 341–348.

[16] D. C. G. Pedronette, O. A. B. Penatti, R. T. Calumby, and R. da S. Torres, "Unsupervised distance learning by reciprocal knn distance for image retrieval," in *International Conference on Multimedia Retrieval (ICMR'14)*, 2014.

[17] 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.

[18] C. J. van Rijsbergen, *Information Retrieval*. London: Butterworth-Heinemann, 1979.

[19] H. P. Young, "An axiomatization of borda's rule," *Journal of Economic Theory*, vol. 9, no. 1, pp. 43–52, 1974.

[20] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009, pp. 758–759.

[21] A. Khudyak Kozorovitsky and O. Kurland, "Cluster-based fusion of retrieved lists," in *ACM SIGIR conference on Research and development in Information Retrieval (SIGIR '11)*, 2011, pp. 893–902.

[22] R. da S. Torres and A. X. Falcão, "Contour Salience Descriptors for Effective Image Retrieval and Analysis," *Image and Vision Computing*, vol. 25, no. 1, pp. 3–13, 2007.

[23] N. Arica and F. T. Y. Vural, "BAS: a perceptual shape descriptor based on the beam angle statistics," *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1627–1639, 2003.

[24] D. C. G. Pedronette and R. da S. Torres, "Shape retrieval using contour features and distance optmization," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP'2010)*, vol. 1, 2010, pp. 197 – 202.

[25] H. Ling and D. W. Jacobs, "Shape classification using the inner-distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 286–299, 2007.

[26] H. Ling, X. Yang, and L. J. Latecki, "Balancing deformability and discriminability for shape matching," in *European Conference on Computer Vision (ECCV'2010)*, vol. 3, 2010, pp. 411–424.

[27] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal on Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[28] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97)*, 1997, pp. 762–768.

[29] R. O. Stehling, M. A. Nascimento, and A. X. Falcão, "A compact and efficient image retrieval approach based on border/interior pixel classification," in *ACM Conference on Information and Knowledge Management (CIKM'2002)*, 2002, pp. 102–109.

[30] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[31] V. Kovalev and S. Volmer, "Color co-occurence descriptors for querying-by-example," in *International Conference on Multimedia Modeling*, 1998, p. 32.

[32] B. Tao and B. W. Dickinson, "Texture recognition and image retrieval using gradient indexing," *Journal of Visual Comunication and Image Representation*, vol. 11, no. 3, pp. 327–342, 2000.

[33] L. J. Latecki, R. Lakmper, and U. Eckhardt, "Shape descriptors for non-rigid shapes with a single closed contour," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2000)*, 2000, pp. 424–429.

[34] J. van de Weijer and C. Schmid, "Coloring local feature extraction," in *European Conference on Computer Vision (ECCV'2006)*, vol. Part II, 2006, pp. 334–348.

[35] P. Brodatz, *Textures: A Photographic Album for Artists and Designers*. Dover, 1966.

[36] B. Li, L.-Y. Duan, Y. Chen, R. Ji, and W. Gao, "Predicting the effectiveness of queries for visual search," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 2361–2364.