Pre-print version of paper published in the Neurocomputing Journal, 2019.
DOI: https://doi.org/10.1016/j.neucom.2019.09.065

Unsupervised Selective Rank Fusion for Image Retrieval Tasks

Lucas Pascotti Valem, Daniel Carlos Guimarães Pedronette

Department of Statistics, Applied Mathematics and Computing (DEMAC), São Paulo State University (UNESP), Rio Claro, Brazil

Abstract

Several visual features have been developed for content-based image retrieval in last decades, including global, local and deep learning based approaches. However, despite the huge advances on features development and mid-level representations, a single visual descriptor is often insufficient to achieve effective retrieval results in several scenarios. Mainly due to the diverse aspects involved in human visual perception, the combination of different features has been establishing as a relevant trend in image retrieval. An intrinsic difficulty consists in the task of selecting the features to combine, which is often supported by supervised learning approaches. Therefore, in the absence of labeled data, selecting features in an unsupervised way is a very challenging, although essential task. In this paper, an unsupervised framework is proposed to select and fuse visual features in order to improve the effectiveness of image retrieval tasks. The framework estimates the effectiveness and correlation among features through a rank-based analysis and use a list of ranker pairs to determine the selected features combinations. High-effective retrieval results were achieved through a comprehensive experimental evaluation conducted on 5 public datasets, involving 41 different features and comparison with other methods. Relative gains up to +55% were obtained in relation to the highest effective isolated feature.

Keywords: content-based image retrieval, unsupervised late fusion, rank-aggregation, correlation measure, effectiveness estimation

1. Introduction

The quick development of visual acquisition technologies and the huge growth of image and multimedia collections have made the use of retrieval and computer vision techniques indispensable. Relevant applications have been proposed in many different associated fields, from text detection in images with complex backgrounds [89] to automatic generation of natural language sentences which summarize the video contents [88]. In this scenario, Content-Based Image Retrieval (CBIR) systems can be broadly defined as any technology which helps to organize images based on their visual content [19]. The most common application consists in retrieving the most similar images to a query image in a given dataset.

Many significant progress have been made in related CBIR areas over the last decades [19, 38, 102]. However, despite the significant recent advances specially in feature extraction methods, effectively retrieving images still remains a challenge in various scenarios. Such complexity is mainly associated to the diverse aspects involved in the human visual perception, which usually can not be encoded by a single visual feature [34, 67]. Images are often composed by complex foregrounds and backgrounds, which makes image understanding techniques a hot research topic and a challenging task [87]. Although deep learning approaches have been producing very significant results, the state-of-art is not reached in all situations [83], turning the selection and combination of different visual features an attractive alternative.

Given the myriad of available visual descriptors, several fusion approaches have been recently proposed [2, 7, 67, 80, 95] aiming at producing more effective retrieval results. Fusion strategies are typically categorized in two different categories: early and late fusion [3, 70], according to the step of the retrieval pipeline where the similarity obtained from different visual features are combined. While early fusion is usually used to combine raw features (e.g. concatenation), late fusion combines different types of representations obtained from the feature vectors (e.g. ranked lists). Recently, relevant unsupervised late fusion methods have been proposed in the literature [101], mainly supported by graph-based approaches [63, 95, 98].

Despite the success of fusion approaches, other crucial task consists in to select what features to combine. In fact, among the great variety of visual features current available, the task of choosing one combination that best fits the need for a given retrieval scenario is a very challenging task [67]. It is known that, finding the optimal combination of ranked lists produced

by different features is a NP-hard problem [22]. Therefore, optimization strategies as genetic programming approaches [17, 24] represent an attractive solution.

Mainly due to the possibility of exploiting the labeled data in order to infer the effectiveness of each visual feature, most of selection methods require training data [9, 67]. In this scenario, various supervised feature selection approaches have been proposed [67]. However, even using supervised learning methods to exploit labeled data, selecting high-effective combinations of visual features remains a complex task, since it is necessary to consider various aspects, as diversity and complementarity of retrieval results.

Therefore, selecting features in an unsupervised way, without any labeled data is even challenger, since no information about effectiveness of individual visual features is available. In several unsupervised late fusion scenarios [64, 95], the selection of visual features is often performed *ad hoc*. Unsupervised feature selection methods [10, 27, 41, 93, 96] based on early fusion strategies [67] can represent an alternative. However, some drawbacks can harm its use in retrieval scenarios, since some methods are very sensible to sparse vectors and they are computationally costly.

In the other hand, unsupervised selection approaches based on late fusion strategies are very scarce in the literature. Some initial selective approaches have been proposed, mainly driven through the task of assigning weights to each visual feature in an unsupervised way [7, 63, 98]. However, there is a lack of unsupervised approaches based on late fusion which explicitly selects the visual features in image retrieval scenarios.

In this paper, we aim at filling this gap. We address this problem by proposing a novel unsupervised framework to select and fuse visual features in order to improve the effectiveness of image retrieval tasks. The proposed framework uses a late fusion strategy supported by rank-based effectiveness estimation measures, which are used to identify the most effective visual features. Additionally, rank correlation measures are used for analyzing complementarity among features. Based on both effectiveness and correlation properties, a list of pairs of visual features is computed and used to determine the selected combination.

The main contributions of the proposed approach in face of the related work are highlighted as follows:

- A completely unsupervised rank-based framework for selecting and fusing visual features in image retrieval scenarios. The selection is per-

formed based on the analysis of top-k retrieval results, requiring low computational costs in comparison with diffusion process [7];

- Different from related late-fusion methods[7, 63, 98], our approach does not assign weights, but explicitly selects the features, favoring the selection in scenarios with larger sets of features;

- The density of reciprocal rank references [59, 65] is exploited to estimate the effectiveness of features and rank correlation measures [53] to encode complementarity information. Since only rank information is required, the framework is flexible, being independent of distance functions and allowing the use of different measures.

A comprehensive experimental evaluation was conducted on five public datasets, involving 41 visual features, among global, local and CNN-based features. The proposed approach was also evaluated in different retrieval scenarios, compared with baselines and other state-of-the-art retrieval methods. High-effective retrieval results were obtained, yielding relative gains up to +55% in relation to the best-isolated feature. The results are comparable or superior to various state-of-the-art retrieval approaches, even when the compared methods use *ad-hoc* pre-selection of features.

The remaining of this paper is organized as follows: Section 2 presents the formulation of the rank retrieval model considered, while Section 3 presents the proposed selection and fusion framework. Section 4 reports the experimental evaluation results. Finally, Section 5 draws the conclusions and possible future work.

2. Rank Model and Problem Setting

This section formally defines the rank model used along the paper. Let $\mathcal{C}=\{x_1, x_2, \ldots, x_N\}$ be an image collection, where N denotes the collection size. Let us consider a retrieval task where, given a query image, 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 $\tau_q=(x_1, x_2, \ldots, x_k)$ in response to the query. The ranked list τ_q can be defined as a permutation of the k-neighborhood set $\mathcal{N}(q)$, 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 $\mathcal{N}(q)$ onto the set $[k] = \{1, 2, \ldots, k\}$. The $\tau_q(i)$ notation denotes the position (or rank) of image x_i in the ranked list τ_q.

The ranker R can be defined based on diverse approaches, including feature extraction or learning methods. In this paper, a feature-based approach is considered, defining R as a tuple (ϵ, ρ), where $\epsilon : \mathcal{C} \to \mathbb{R}^d$ is a function that extracts a feature vector v_x from an image $x \in \mathcal{C}$; and d: $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a distance function that computes the distance between two images according to their corresponding feature vectors. Formally, the distance between two images x_i, x_j is defined by $d(\epsilon(x_i), \epsilon(x_j))$. The notation $d(x_i, x_j)$ is used for readability purposes.

A ranked list can be computed by sorting images in a crescent order of distance. In terms of ranking positions we can say that, if image x_i is ranked before image x_j in the ranked list of image x_q, that is, $\tau_q(i) < \tau_q(j)$, then $d(q, i) \leq d(q, j)$. Taking every image in the collection as a query image x_q, a set of ranked lists $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_n\}$ can be obtained.

Different features and distance functions give rises to different rankers which, in turn, produce distinct ranked lists. Let $\mathcal{R} = \{R_1, R_2, \ldots, R_m\}$ be a set of rankers and $R_j \in \mathcal{R}$, we denote by $\mathcal{T}_j$ the set of ranked lists produced by R_j. A ranked list computed by the ranker R_j in response to a query x_q is denoted by $\tau_{j,q}$.

The objective of the proposed selection framework is to select from the set $\mathcal{R}$ the rankers which produces the most effective retrieval results, based on their respective set of ranked lists, without the need of any labeled data. Formally, the framework can be defined by a set function f_s as follows:

$$\mathfrak{X}_n^* = f_s(\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_m), \tag{1}$$

where $\mathfrak{X}_n^*$ denotes the set of rankers selected by the framework for a given size n, such that $|\mathfrak{X}_n^*| = n$

3. Unsupervised Selective Rank Fusion

This section describes the proposed Unsupervised Selective Rank Fusion (USRF) method. The main motivation of the proposed approach consists in to select and combine a set of rankers in a completely unsupervised way, considering scenarios where several visual features are available and there are no labeled data or feedback from the user. Figure 1 presents a diagram which illustrates the use of the USRF method, in terms of its input and

output data. Given an image dataset, it shows that rankers based on different approaches can be used (color descriptors, bag of words representations, convolutional neural networks, and others) with the intent of exploiting the complementarity of the data and producing more effective retrieval results.

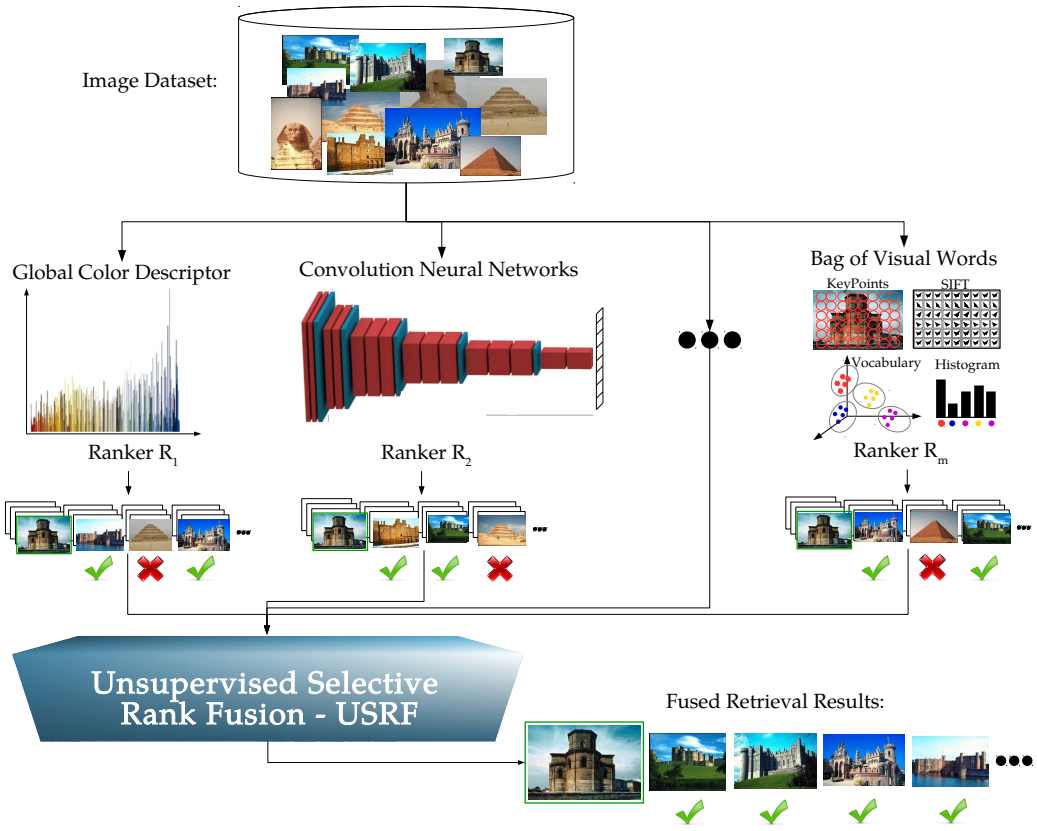


Figure 1: General view of the use of USRF method.

Let $\mathfrak{R} = \{R_1, R_2, ..., R_m\}$ denotes a set of m different rankers. The Cartesian production $\mathfrak{R}^2 = \mathfrak{R} \times \mathfrak{R}$ defines a set of all of the possible unordered pairs $\{R_i, R_j\} \in \mathfrak{R}^2$ of rankers. In this work, we consider not only pairs, but sets of combinations composed of two or more rankers. Therefore, we can generalize the set of ranker combinations for different sizes as:

$$\mathfrak{R}^n = \prod_{1}^{n} \mathfrak{R}, \tag{2}$$

which contains all the combinations with size n. The full set of possible combinations available for the USRF selection is given by the union of all combinations with different sizes. Formally, we can define the search space represent by the full set of combinations $\mathcal{S}$ as:

$$\mathcal{S} = \cup_{i=1}^{m} \mathfrak{R}^i. \tag{3}$$

For the ease of the reader, Table 1 summarizes the symbols used through the definition of our approach (some of them will be discussed in details in next sub-sections).

Table 1: Table of symbols

Type	Symbol	Description
Retrieval Model	$\mathcal{C}$	Image collection.
	$\mathcal{N}(q,k)$	Neighborhood set for a query image q of size k.
	τ_q	Ranked list for the query image q.
	$\tau_q(j)$	Position of the image j in the ranked list of the image q.
	L	Size of the ranked lists.
	$\mathcal{T}$	Set of ranked lists for all the images in the dataset.
Selection Model	f_s	Function for ranker selection.
	R_i	Ranker of index i.
	$\tau_{i,q}$	Ranked list of the image q computed by the ranker i.
	$\mathcal{T}_i$	Set of ranked lists produced by the ranker R_i.
	$\mathfrak{R}$	Set of rankers used as input for the USRF.
	m	Size of the set $\mathfrak{R}$.
	$\mathcal{S}$	Selection set for all the available combinations of rankers.
	$\mathfrak{X}_n$	Candidates of rankers combination of size n.
	$\mathfrak{X}_n^*$	Combination of size n selected by the USRF.
	C_n	Set of combinations where each combination is of size n.
	τ_n^R	List of all the combinations of size n sorted by the selection measure.
	$\tau_n^R(\mathfrak{X}_n^i)$	Position of the combination $\mathfrak{X}_n^i$ in τ_n^R.
	L_R	Size of the ranked combinations list τ_n^R.

For a given search space $\mathcal{S}$, and combination size n, we aim at selection a combination $\mathfrak{X}_n^*$. In such **unsupervised** scenario, the most challenging aspects associated to the selection of visual features can be summarized as:

(i) **Computation cost:** the computational cost to evaluate all the combinations is often prohibitive;

(ii) **Lack of Information:** there is absolutely no information about the quality of visual features. Therefore, it is not possible to learn a selection function capable of evaluate rankers based on training data.

Such challenges are addressed by the proposed USRF method based on two key ideas:

(i) **Selection Based on List of Pairs:** while the set of all the combinations $\mathcal{S}$ is huge (even for a small input size $|\mathfrak{R}|$), the number of combinations given by unordered pairs $\mathfrak{R}^2$ grow much more slowly. Based on this observation, we derive a broadly selection combination based on a selection of pairs or rankers, drastically reducing the computational efforts required.

(ii) **Unsupervised Selection Measure:** since our approach performs the selection based on pairs, an unsupervised selection measure is proposed in order to identify the most promising ranker pairs to be combined. Firstly, the proposed measure exploits unsupervised effectiveness measures to obtain an approximation of quality of visual features. Such measures are based on the density of ranking references at top positions of retrieval results, which requires no training data. In addition, rank correlation measures are also computed for each pair, evaluating similarity and complementarity among retrieval results.

3.1. Overall Organization of USRF

Figure 2 illustrates the overal organization of USRF, highlithing its main steps (1 to 4) and the sequence in which they occur. The method input consists in sets of ranked lists ($\mathfrak{R}$) computed by different visual features (showed at the top of the figure).

In step (1), an effectiveness estimation is computed for each ranker, while in step (2) the correlation measure is computed for each pair of rankers. Subsequently, a measure is applied with the objective of selecting the best combination (3), according to the values obtained in (1) and (2). Notice that the selected combination is denoted by $\mathfrak{X}^*$. Finally, in step (4), the selected rankers are offered as input for the rank-aggregation method, which computes the final retrieval results. Next subsections discusses each of the steps in details.

3.2. Unsupervised Effectiveness Estimation

Retrieval approaches are often evaluated based on a quantitative value given by effectiveness measures (as Precision, Recall, MAP, NDCG), which are computed based on labeled data. In the unsupervised scenario addressed in this paper, we propose to exploit effectiveness estimation measures [62, 74, 86], which does not required labeled data.

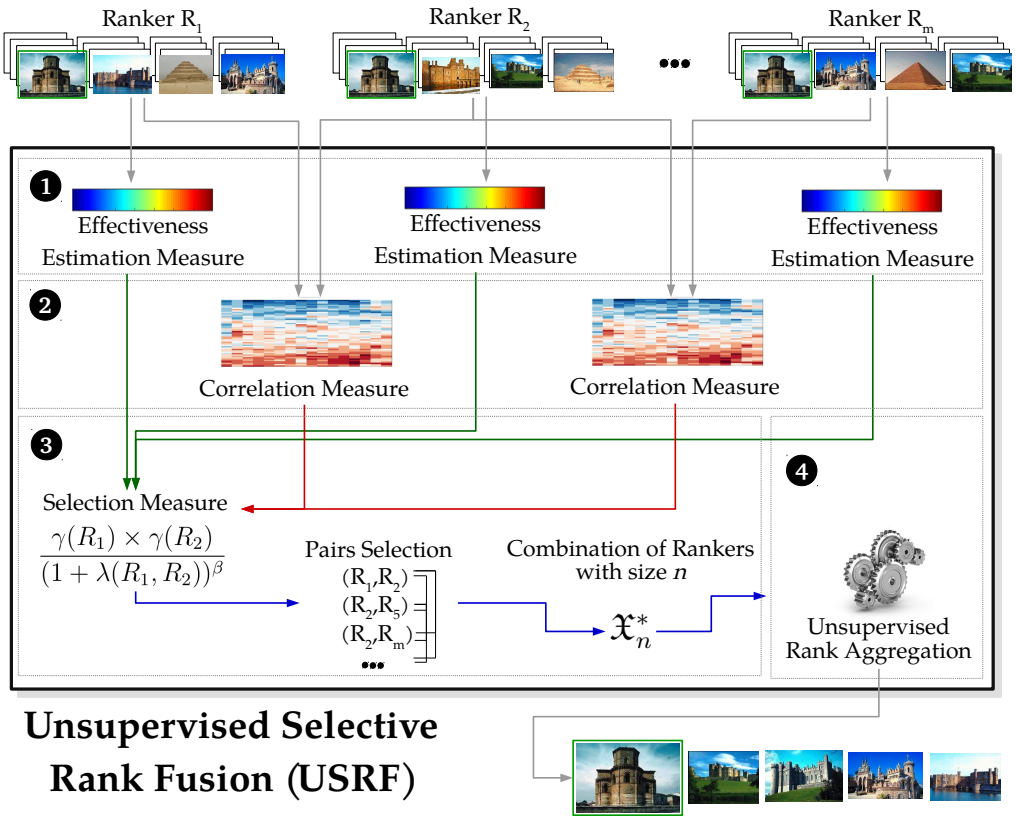


Figure 2: Illustration of the steps that compose the USRF.

In general, such measures analyzes contextual information, as the relationship among ranking references, in order to provide an estimation of effectiveness of each ranked list. A real value in the interval $[0, 1]$ is assigned to each ranked list in the dataset, which is used for providing an effectiveness estimation of the ranker. Next, we present the effectiveness estimation measures used in this work.

- ***Authority Score***

The Authority Score [65] is a graph-based effectiveness estimation measure. The measure uses a graph representing the ranking references at top positions of ranked lists and estimates the effectiveness according to the density of the graph. Each image in the top-k positions of the ranked list τ_q defines a node. For each image x_j in the top-k of τ_q, the ranked list τ_j is also

analyzed. If there are images in common in ranked lists τ_q and τ_j, an edge is created. The authority score is computed based on the number of created edges. Therefore, the measure is based on the density of the graph and can be formally defined as follows:

$$Authority(\tau_q, k) = \frac{\sum_{u \in \mathcal{N}(q,k)} \sum_{v \in \mathcal{N}(u,k)} f_{in}(v, q)}{k^2}, \quad (4)$$

where f_{in} returns 1 if $\tau_q(v) \leq k$ and 0 otherwise. The Authority Score is defined in the interval $[0, 1]$, achieving the greater score for a full connected graph at top-k positions.

- ***Reciprocal Density***

  The Reciprocal Density [59] also exploits the ranking references considering the density of reciprocal neighbors at top-k positions. A weight is assigned to the occurrence of each reciprocal neighbor according to its position in the ranked lists. The formal definition is given by the Equation 5.

$$Reciprocal(\tau_q, k) = \frac{\sum_{i \in \mathcal{N}(q,k)} \sum_{j \in \mathcal{N}(i,k)} f_{in}(j, q) \times w_r(q, i) \times w_r(i, j)}{k^4}, \quad (5)$$

  where a weight is computed to each position according to the function $w_r(q, i) = k + 1 - \tau_q(i)$. The higher the weight, higher tends to be the occurrence of reciprocal neighbors in the first positions of the ranked lists.

3.3. Rank Correlation Measures

A retrieval task can benefit from a fusion approach if the inputs being combined contain diverse and complementary information. Such complementarity can be analyzed directly based on ranking information, since it is expected that the top positions of rankings produced by different rankers contain distinct relevant results.

- **Jaccard**

  The Jaccard index is a statistic measure that computes the correlation between two ranked lists based on its intersection and is defined by Equation 6:

$$J(\tau_i, \tau_j, k) = \frac{|\mathcal{N}(u, k) \cap \mathcal{N}(v, k)|}{|\mathcal{N}(u, k) \cup \mathcal{N}(v, k)|}, \quad (6)$$

where $\mathcal{N}(u)$ and $\mathcal{N}(v)$ denote the k-neighborhood sets which contain the elements of the ranked lists τ_u and τ_v, respectively.

- **Jaccard$_k$**

  The traditional Jaccard coefficient performs its analysis at a single depth defined by k. In this way, the same weight is assigned to all objects at top-k positions. In [53], a Jaccard score considering different depths is proposed, assigning higher weights to top positions. This measure is defined by:

$$J_k(\tau_i, \tau_j, k) = \frac{\sum_{d=1}^{k} J(\tau_i, \tau_j, d)}{k}. \tag{7}$$

- **RBO**

  The Rank-Biased Overlap (RBO) [82] also considers the overlap between top-k lists at increasing depths. However, different from the intersection measure, the weight of the overlap measure is computed based on probabilities defined at each depth. The RBO measure is defined by:

$$RBO(\tau_i, \tau_j, k, p) = (1-p) \sum_{d=1}^{k} p^{d-1} \times \frac{|\mathcal{N}(i,k) \cap \mathcal{N}(j,k)|}{d}, \tag{8}$$

  where $p = 0.9$ was considered for all the experiments.

- **Spearman**

  The Spearman's metric is a non-parametric measure, which evaluates the relationship between two variables. Usually denoted by the letter ρ, it can be seen as the L1 distance between two permutations, considering the difference between positions of elements [23]. The measure is formally defined as:

$$Spearman(\tau_i, \tau_j, k) = 1 - \frac{\sum_{x,y \in \mathcal{N}(i,k) \cup \mathcal{N}(j,k)} s(x,y)}{2 \times k^2}, \tag{9}$$

  where $s(x,y)$ is the difference between positions computed by:

$$s(x,y) = |max(k+1, \tau_x(y)) - max(k+1, \tau_y(x))|. \tag{10}$$

  Notice that we propose to restrict the difference to k positions due to the pre-processing given by the neighborhood sets.

- **Kendallτ**

  The Kendall's τ is a traditional correlation measure between permutations, computed based on the number of exchanges needed in a bubble sort to convert one permutation to the other [23]. The measure is commonly used as a rank correlation measure and can be defined as follows:

$$Kendall_{\tau}(\tau_i, \tau_j, k) = 1 - \frac{\sum_{x,y \in \mathcal{N}(i,k) \cup \mathcal{N}(j,k)} \bar{K}_{x,y}(\tau_i, \tau_j)}{k \times (k-1)}, \tag{11}$$

  where $\bar{K}_{x,y}(\tau_i, \tau_j)$ is a function that determines if objects o_x and o_y are in the same order in top-k lists τ_i and τ_j. Formally, the function can be defined as follows:

$$\bar{K}_{x,y}(\tau_i, \tau_j) = \begin{cases} 0 & \text{if } (\tau_i(x) \leqslant \tau_i(y) \wedge \tau_j(x) \leqslant \tau_j(y)), \\ 0 & \text{if } (\tau_i(x) \geqslant \tau_i(y) \wedge \tau_j(x) \geqslant \tau_j(y)), \\ 1 & \text{otherwise.} \end{cases} \tag{12}$$

  For the computation of $\bar{K}$ function, we considered the maximum position as k (positions bigger than k are set to $k+1$).

3.4. Selection Strategy

As previously stated, given a set $\mathfrak{R}$ composed by m rankers, USRF searches for the best combinations in the set $\mathcal{S} = \cup_{i=1}^{m} \mathfrak{R}^i$. We refer to a combination as a set of two or more rankers, which is denoted by $\mathfrak{X}_n^i$, where i indicates the index and n the size of the combination, respectively.

Section 3.4.1 describes the proposed selection approach for ranker pairs (conducted on the set $\mathfrak{R}^2$), while Section 3.4.2 presents the selection for combinations of different sizes (conducted on the set $\mathcal{S}$) and how it is derived from the selection of pairs.

3.4.1. Selection of Ranker Pairs

An unsupervised selection measure is proposed, assigning a score for each pair of rankers through a function w_p. The selection measure is based on two different hypotheses:

1. The higher the effectiveness estimation of a ranker, higher the chances of it offering an effective result when combined with the others;

2. The lower the correlation between two rankers, higher the chances of complementary information, which can be combined to obtain a more effective result.

Such hypotheses were already exploited in previous work [77], although requiring labeled data and traditional. In the following, we discuss the selection performed based on effectiveness, correlation or a combination of them.

- **Selection through Effectiveness Estimation**

  The first hypothesis is based on the idea that a relevant combination is given by ranked lists of high effectiveness. Let Γ be the effectiveness selection measure applied to a pair $\{R_1, R_2\}$ and $\gamma(R_i)$ be a function that returns the effectiveness estimation of ranked lists offered by the ranker R_i. The measure is defined as:

$$\Gamma(R_1, R_2) = \gamma(R_1) \times \gamma(R_2). \tag{13}$$

  Notice that $\gamma(R_i)$ can refer to any of the measures presented in Section 3.2.

- **Selection through Correlation**

  Another hypothesis is that ranked lists with low correlation provide a potencial way for exploiting the complementarity of the data. Let Λ be the correlation selection measure and $\lambda(R_1, R_2)$ be a function that returns a value of correlation (similarity in the interval $[0, 1]$) between the ranked lists offered by the rankers R_1 and R_2. The measure is defined by the Equation 14.

$$\Lambda(R_1, R_2) = \frac{1}{1 + \lambda(R_1, R_2)}. \tag{14}$$

  The function $\lambda(R_1, R_2)$ can consider as correlation measure any of the measures presented in Section 3.3 (Jaccard, RBO, Spearman or Kendallτ, for example).

- **Joint Selection Measure**

  The selection measure for pairs of rankers is proposed based on the incorporation of the two earlier equations, with the objective of selecting pairs of high effectiveness and low correlation (high complementarity). The measure

w_p is defined as:

$$w_p(\{R_1, R_2\}) = \Gamma(R_1, R_2) \times \Lambda(R_1, R_2)^\beta = \frac{\gamma(R_1) \times \gamma(R_2)}{(1 + \lambda(R_1, R_2))^\beta}, \tag{15}$$

where the exponent β is used with the intent of applying a weight for the correlation, which provides the use of the proposed measure in different scenarios. Along the experiments, we noticed that for the cases where there are a high number of rankers, it is more beneficial to use the correlation to combine ranked lists that are similar. Therefore, $\beta = 1$ was adopted for scenarios with less diversity and $\beta = -1$ was adopted for scenarios with higher dversity, as discussed in more details in the experimental evaluation (Section 4).

- **Pairs Selection**

  Finally, the ranker pairs can be sorted in a decreasing order of selection function w_p, which can be used to obtain the ranked lists of the pairs denoted by τ_2^R. More formally, a ranked pairs list $\tau_2^R = (\mathfrak{X}_2^1, \mathfrak{X}_2^2, ..., \mathfrak{X}_2^{L_R})$ can be defined as a permutation of $\mathfrak{R}^2$. The permutation τ_2^R is the bijection of the set $\mathfrak{R}^2$ in $[L_R] = \{1, 2, ..., L_R\}$, where L_R is a parameter which defines the maximum size of the permutation. For a permutation τ_2^R, we consider $\tau_2^R(\mathfrak{X}_2^i)$ as the position of $\mathfrak{X}_2^i$ in τ_2^R.

  We can say that if $\tau_2^R(\mathfrak{X}_2^i) \leq \tau_2^R(\mathfrak{X}_2^j)$ then $w_p(\mathfrak{X}_2^i) \geq w_p(\mathfrak{X}_2^j)$. Therefore, the element in the first position is the most effective and so on. The selected pair can be defined by the equation:

$$\mathfrak{X}_2^* = \underset{\mathfrak{X}_2^i \in \tau_2^R}{\arg\max}\, w_p(\mathfrak{X}_2^i). \tag{16}$$

3.4.2. Selection of Rankers Set

Based on the selection of pairs, this section extends the approach for selecting combinations of any number of rankers. The method applies the selection for combinations of any size through the union of the most relevant pairs, which is indispensable to guarantee the selection for larger rankers sets.

Figure 3 illustrates the algorithm for selection of rankers combination. The input is given by the ranked list τ_2^R, which is composed by ranker pairs sorted according to w_p. The intersection among pairs $\{R_2, R_3\}$, $\{R_1, R_2\}$, $\{R_1, R_3\}$ gives rise to the combination $\{R_1, R_2, R_3\}$, which appear in τ_3^R.

The value of w corresponds to the sum of w_p from the originating pairs. Notice that the pair $\{R_4, R_5\}$ does not lead to any combination in C_3, once this pair does not have any intersection in τ_2^R. The intersection-based process is repeated for obtaining the combinations of four elements (C_4), and can be iteratively repeated for bigger combinations.

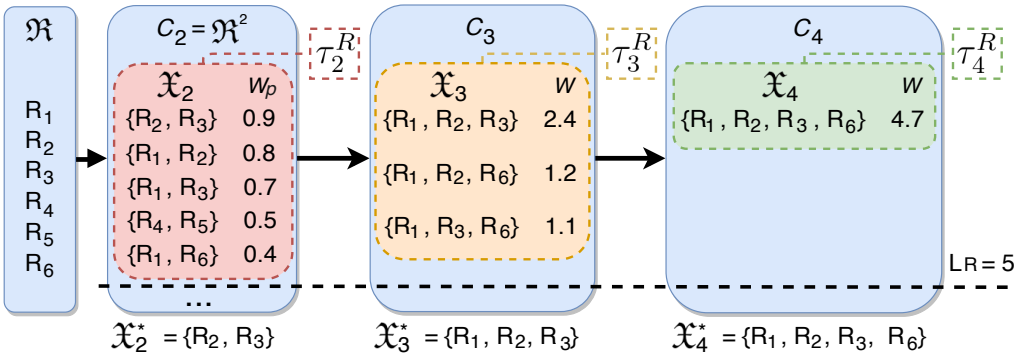


Figure 3: Illustration of the proposed selection method.

The set C_n contains all the combinations of size n formed from the union of the combinations of size $(n-1)$ that belong to τ_{n-1}^R. Formally, the set C_n is defined[1] by the Equation 17.

$$C_n = \begin{cases} \mathfrak{R}^2 & \text{se } n = 2 \\ \{\mathfrak{X}_{n-1}^i, \mathfrak{X}_{n-1}^j \in \tau_{n-1}^R \ \wedge \ |\mathfrak{X}_{n-1}^i \cup \mathfrak{X}_{n-1}^j| = n\} & \text{se } n \geq 3 \end{cases} \tag{17}$$

As the function w_p defines a score only for pairs of rankers, it can be generalized for combinations of any size through a function w, defined by the Equation 18. It applies a sum of the values of w from the combinations of size $(n-1)$ recursively, which has as base case the value obtained in w_p.

$$w(\mathfrak{X}_n) = \begin{cases} w_p(\mathfrak{X}_n) & \text{se } n = 2 \\ \sum_{\mathfrak{X}_{n-1}^i \in \tau_{n-1}^R} w(\mathfrak{X}_{n-1}^i) & \text{se } n \geq 3 \end{cases} \tag{18}$$

The value of w can be computed for all the combinations in the collection C_n. In the following the combinations can be sorted in a decreasing of w to obtain a ranked list of the combinations τ_n^R.

[1]According to the union operation for a given size, the set C_n can be empty.

Formally, the ranked list of ranker combinations $\tau_n^R = (\mathfrak{X}_n^1, \mathfrak{X}_n^2, ..., \mathfrak{X}_n^{L_R})$ can be defined as a permutation of $C_n \subset \mathfrak{R}^n$. The permutation τ_n^R is a bijection of the set C_n in $[L_R] = \{1, 2, ..., L_R\}$, where L_R is defines the size of the list.

Therefore, generalizing the selection algorithm, the process for obtaining the selected combination $(\mathfrak{X}_n^*)$ of any size (n) is defined by the Equation 19. Alternatively, the selected combination $\mathfrak{X}_n^*$ can also be understood as the one ranked in the first position of ranked combinations list τ_n^R.

$$\mathfrak{X}_n^* = \underset{\mathfrak{X}_n^i \in \tau_n^R}{\arg\max}\, w(\mathfrak{X}_n^i). \tag{19}$$

Algorithm 1 presents the pseudo-code for an efficient algorithmic solution for the proposed method. The input is given by a set of rankers $\mathfrak{R}$ and the size t of the combination to be selected. The presented approach still does not offer a strategy to automatically define the size of the combination to be selected, which is one of the possibilities for future work.

The selection process is based on the selection of pairs, which occurs in the lines (1) and (2). While in (1) the set C_2 is initialized with all of the available pairs obtained from $\mathfrak{R}$, in (2) a ranked list of the pairs is obtained. The function $sort(C_n, w, L_R)$ returns a list which contains the first top-L_R combinations of C_n decreasingly sorted by the function w which is internally computed by the function *sort*. Between (3) and (12) the pseudo-code describes the iterative process for obtaining the sets C_n and τ_n^R for different values of n until reaching the value of t. We can say that the Equation 17 is equivalent to the process described by the lines (1), for pairs, and (4) to (10) for combinations of any size. Finally, (13) describes the process for obtaining the selected combination $\mathfrak{X}_t^*$ according to the Equation 19.

3.5. Aggregation Method

Given a selected combination of rankers, the objective consists in to aggregate the retrieval results computed by each ranker. The USRF is very flexible and can use different aggregation methods for performing the aggregation of rankers. In this work, we use the CPRR (Cartesian Product of Ranking References) [76] method, a recent proposed unsupervised approach which presents effective and efficient results comparable to the state-of-the-art. The CPRR has as its central idea the use of the Cartesian product with the objective of maximizing the contextual information coded in the

Algorithm 1 Selection of Rankers Combination

Require: Set of rankers $\mathfrak{R}$ and the size t of the combination to be selected.
Ensure: Selected combination $\mathfrak{X}_t^*$.

$C_2 \leftarrow \mathfrak{R}^2$
$\tau_2^R \leftarrow sort(C_2, w, L_R)$
for $n = 3$ to t **do**
    $C_n \leftarrow \emptyset$
    for all $\mathfrak{X}^i, \mathfrak{X}^j \in \tau_{n-1}^R$ **do**
        $\mathfrak{X}^u \leftarrow \mathfrak{X}^i \cup \mathfrak{X}^j$
        if $|\mathfrak{X}^u| = n$ **then**
            $C_n \leftarrow C_n \cup \{\mathfrak{X}^u\}$
        end if
    end for
    $\tau^R \leftarrow sort(C_n, w, L_R)$
end for
$\mathfrak{X}_t^* \leftarrow \tau_t^R(1)$
return $\mathfrak{X}_t^*$

ranked lists. For efficiency and scalability reasons, only the subset of the top-L images of the ranked lists is considered.

4. Experimental Evaluation

This section presents the experimental evaluation conducted in order to asses the effectiveness of the proposed approach. Section 4.1 discusses the experimental protocol, describing datasets, evaluation measures, visual features, and the methods considered as baselines. Section 4.2 analyzes the proposed method considering different aspects, including parameters and measures. Section 4.3 presents the obtained results in terms of selection and combination tasks, while Section 4.4 compares the obtained results to other state-of-the-art methods. Finally, Section 4.5 shows some visual results.

4.1. Experimental Protocol

The experimental protocol is presented as follows: Section 4.1.1 describes the datasets; Section 4.1.2 discusses the visual features and Section 4.1.3 the selection scenarios considered; and Section 4.1.4 presents the methods used as baselines.

4.1.1. Datasets

The experimental analysis considered five different datasets with sizes ranging from 1,360 to 10,200 images, which are presented in Table 2. Most of them are frequently used for image retrieval tasks. The MAP was computed considering every image in the dataset as a query, except for the Holidays dataset where a specific set of queries was considered following the protocol proposed by the authors of the dataset in [31].

Besides the MAP, Recall@40 and N-S Score were used for the datasets MPEG-7 and UKBench, respectively. These measures were employed to facilitate the comparison of our results with state-of-the-art baselines.

Table 2: Datasets considered in the experimental evaluation.

Dataset	**Size**	**Type**	**General Description**	**Effectiv. Measure**
Flowers [50]	1,360	Flowers	Composed of 17 species of flowers with 80 images of each presenting pose and light variations. This dataset is distributed by the University of Oxford.	MAP
MPEG-7 [37]	1,400	Shape	Composed of 1,400 shapes divided in 70 classes. Commonly used for evaluation of post-processing methods.	MAP, Recall@40
Holidays [31]	1,491	Scenes	Commonly used as image retrieval benchmark, the dataset is composed of 1,491 personal holiday pictures with 500 queries.	MAP
Corel5k [44]	5,000	Objects/ Scenes	Composed of 50 categories with 100 images for each class, including diverse scene content such as fireworks, bark, microscopy images, tiles, trees, etc.	MAP
UKBench [51]	10,200	Objects/ Scenes	Composed of 2,550 objects or scenes. Each object/scene is captured 4 times from different viewpoints, distances, and illumination conditions.	MAP, N-S Score

4.1.2. Visual Features

A wide variety of visual features were considered, including different categories: global, local, and deep learning. Table 3 presents each of them followed by their respective types, references, a short description, and the MAP obtained for each dataset. Notice that, for all the datasets, the set of applied descriptors is similar, except for the MPEG-7 dataset which presents a more specific retrieval scenario and, therefore, only shape descriptors were employed.

The deep learning results were obtained using PyTorch [56], one of the most popular open-source frameworks for machine learning. All of the networks were trained on ImageNet [20], a dataset commonly used for training

general purpose convolutional neural networks. The distances were extracted from the feature vectors obtained from the last layer before the classification layer. For the CNN-OLDFP, the features vectors are offered by [69], the network mixes techniques of deep learning and bag of words aiming at optimizing the effectiveness of the results.

Table 3: Information about the descriptors used in the experimental evaluation.

Category	**Type**	**Descriptor**	**Short Description**	**Original MAP (%)**				
				MPEG-7	Flowers	Corel5k	UKBench	Holidays
Global	Color	ACC [29]	Auto Color Correlogram	—	18.99	23.44	87.72	64.29
		SPACC [29, 48]	Spatial Pyramid ACC	—	19.20	23.86	85.30	62.37
		CLD [15]	Color Layout Descriptor	—	18.54	17.86	59.58	37.59
		SCD [15]	Scalable Color Descriptor	—	10.25	14.56	83.04	54.26
		SCH [15]	Simple Color Histogram	—	13.43	17.56	48.98	24.19
		FOH [48, 78]	Fuzzy Opponent Histogram	—	11.42	15.87	57.05	25.77
		BIC [71]	Border/Interior Pixel Classification	—	25.56	—	80.46	—
	Shape	PHOG [18, 48]	Pyramidal Histogram of oriented gradients	—	14.74	15.80	41.60	31.15
		AIR [25]	Articulation-Invariant Representation	89.39	—	—	—	—
		ASC [43]	Aspect Shape Context	85.28	—	—	—	—
		IDSC [42]	Inner Distance Shape Context	81.70	—	—	—	—
		CFD [60]	Contour Features Descriptor	80.71	—	—	—	—
		BAS [1]	Beam Angle Statistics	71.42	—	—	—	—
		SS [16]	Segment Saliences	37.82	—	—	—	—
	Texture	LBP [52]	Local Binary Patterns	—	10.34	14.83	47.19	28.82
		SPLBP [48, 52]	Spatial Pyramid LBP	—	10.92	15.41	52.14	33.09
		EHD [49]	Edge Histogram Descriptor	—	12.46	16.80	44.10	25.83
	Color and Texture	CEDD [11]	Color and Edge Directivity Descriptor	—	20.48	23.00	70.45	51.59
		SPCEDD [11, 48]	Spatial Pyramid CEDD	—	21.94	28.70	74.98	56.09
		FCTH [12]	Fuzzy Color and Texture Histogram	—	20.56	23.93	73.70	48.44
		SPFCTH [12, 48]	Spatial Pyramid FCTH	—	21.73	26.43	77.78	55.43
		JCD [94]	Joint Composite Descriptor	—	20.89	24.73	74.85	52.84
		SPJCD [48, 94]	Spatial Pyramid JCD	—	22.56	28.02	76.67	56.58
		COMO [79]	Compact Composite Moment-Based Descriptor	—	21.83	21.05	79.77	49.66
	Holistic	GIST [54]	Global Image Descriptor for low-dim. features	—	9.82	15.98	45.44	21.59
Local	Bag of Words	SIFT [47]	Scale-Invariant Feature Transform with VLAD	—	28.47	12.60	74.52	54.63
		VOC [81]	Vocabulary Tree	—	—	—	91.14	—
Deep Learning	Convolutional Neural Networks trained on Imagenet	CNN-SENet [28]	154-layers Squeeze-and-Excitation Neural Network	—	43.16	56.92	92.15	71.60
		CNN-ResNet [26]	152-layers Residual Neural Network	—	51.83	64.81	94.54	74.88
		CNN-FBResNet [26]	152-layers ResNet trained by Facebook AI Research	—	52.56	64.21	93.88	72.65
		CNN-ResNeXt [85]	101-layers "Next Generation" ResNet	—	51.91	62.39	93.67	74.16
		CNN-DPNet [13]	92-layers Dual Path Neural Network	—	50.93	65.15	90.47	70.59
		CNN-VGGNet [45]	19-layers VGG Neural Network	—	39.05	47.85	87.99	67.96
		CNN-BnVGGNet [45]	19-layers Binaural VGG Neural Network	—	41.87	52.72	89.24	67.60
		CNN-InceptionV4 [73]	Fourth version of the Inception Neural Network	—	42.35	58.66	86.82	63.84
		CNN-InceptionResNet [73]	Inception architecture with residual connections	—	42.20	61.17	87.23	62.87
		CNN-BnInception [30]	Binaural Inception Neural Network	—	46.58	46.60	91.84	70.06
		CNN-NASnet-Large [103]	Convolutional Neural Architecture Search Network	—	40.74	53.55	86.90	64.48
		CNN-AlexNet [36]	Alex Krizhevesky Convolutional Neural Network	—	46.04	37.67	85.57	65.25
		CNN-Xception [14]	Depthwise Separable Convolutions Neural Network	—	47.31	54.44	90.83	64.94
	Pooled CNN	CNN-OLDFP [69]	Object Level Deep Feature Pooling	—	—	—	97.74	88.46

4.1.3. Selection Scenarios

Along the experimental evaluation, different selection scenarios were considered:

(i) the **full** that considers all the presented descriptors;
(ii) only **global and local** descriptors;
(iii) only **deep learning**;

(iv) the **custom** scenarios, which consider a set of six different descriptors for each dataset.

The custom scenario is composed by the two most effective deep learning descriptors, the most effective local descriptor, and the three most effective global descriptors of different types. Table 4 presents the custom scenarios for each one of the datasets, the descriptors are presented in decreasing MAP order.

Table 4: Descriptors considered in the custom scenarios.

Dataset	**Descriptors**
MPEG-7	AIR, ASC, IDSC, CFD, BAS, SS
Flowers	CNN-FBResNet, CNN-ResNeXt, SIFT, BIC, SPJCD, PHOG
Corel5k	CNN-DPNet, CNN-ResNet, SIFT, SPACC, SPCEDD, EHD
UKBench	CNN-OLDFP, CNN-ResNet, VOC, ACC, COMO, SPLBP
Holidays	CNN-OLDFP, CNN-ResNet, SIFT, ACC, SPJCD, SPLBP

4.1.4. Baselines

With the purpose of presenting a comprehensive experimental analysis, different approaches were considered as baselines. To make a fair comparison with our method, all of the baselines are completely unsupervised. All of them are recent, present results comparable to the state-of-art, and are open-source. This allows us to use the same input for the proposed method and the baselines. Table 5 presents the late fusion methods considered as baselines.

Table 5: Late fusion methods considered as baselines.

Method	**General Description**
Correlation Graph **[63]**	The method exploits the intrinsic geometry of the dataset aiming at defining a more effective distance between images. Among the different employed strategies, the method builds a graph and analyzes its strongly connected components.
Query-Adaptive Fusion **[98]**	Given a matrix that offers the similarity among the images in a dataset, the effectiveness of a descriptor is estimated as inversely proportional to the area below the similarity curve for each element in the collection. The estimations are used to compute more effective similarity values, which are offered as the output of the method.
Graph Fusion **[95]**	For each of the input data, an undirected weighted graph is built considering each one of the images as a query. The graps are fused using different techniques, including the PageRank [55] algorithm.

In addition, five feature selection algorithms based on early approaches were also considered as baselines:

- *Laplacian Score* [27];
- *Spectral Regression* (SPEC) [96];
- *Muti-cluster Feature Selection* (MCFS) [10];
- *Unsupervised Discriminative Feature Selection* (UDFS) [93];
- *Nonnegative Discriminative Feature Selection* (NDFS) [41].

The first two are based on similarity measures and the others are based on the processing of sparse matrices. All of them are publicly available in the *scikit-feature* [39] [2], a python library for feature selection.

4.2. Experimental Analysis

This section presents various experiments which analyzes the proposed framework in different aspects. Section 4.2.1 evaluates the impact of parameters on the retrieval results. Section 4.2.2 assess the proposed method in terms of the effectiveness estimation and rank correlation measures. Section 4.2.3 analyzes the influence of correlation on the selection measure and Section 4.2.4 evaluates the effect of the list size of combinations.

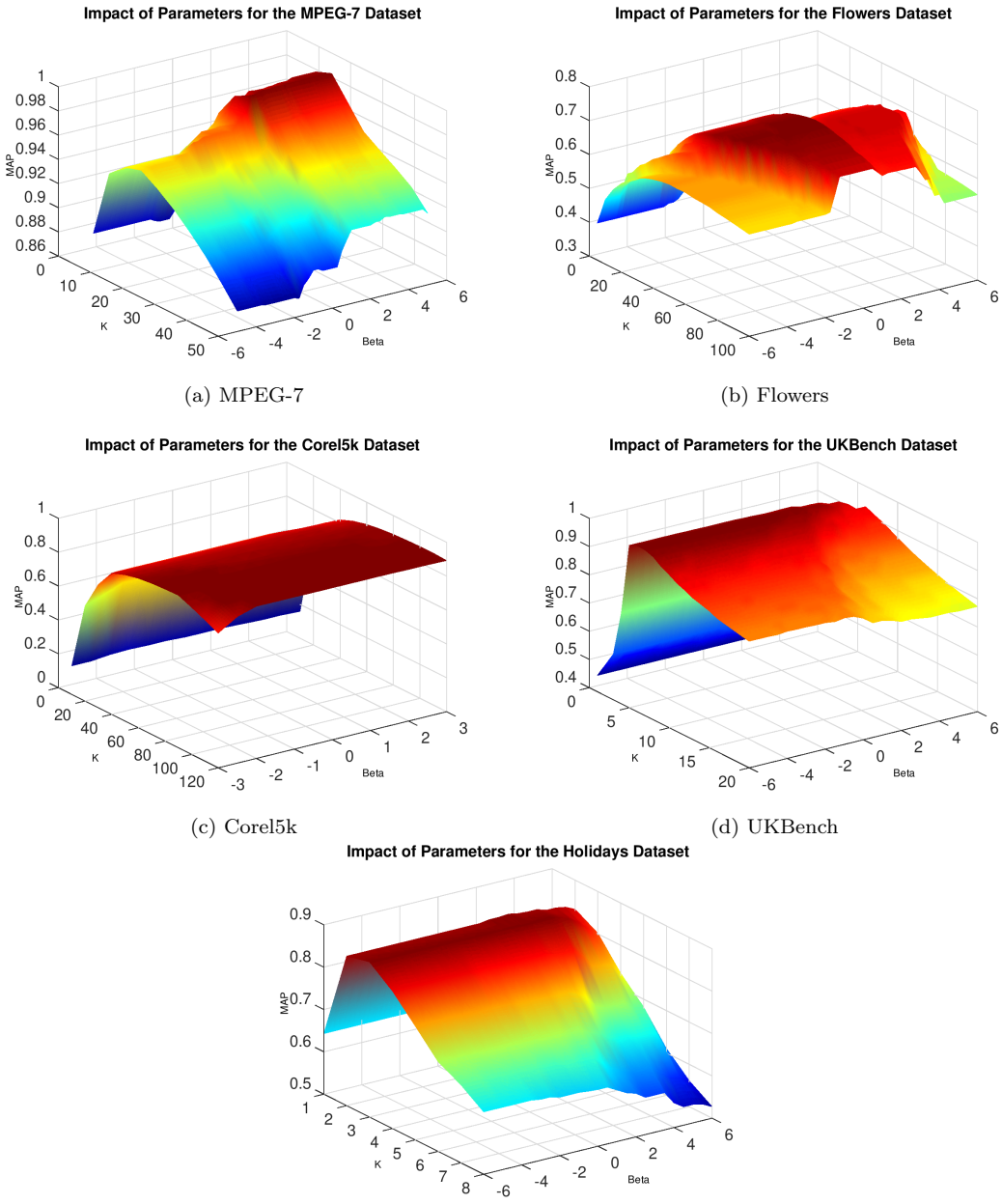
Most of the results presented in this section consider the selection of pairs. For comparison purposes, a weighted arithmetic mean of the MAP is computed for the top-5 pairs ranked by the USRF (weight 5 for the first pair, 4 for the second, and so on).

4.2.1. Impact of parameters

This section analyzes the impact of parameters given neighborhood size k and weight of correlation β. The neighborhood size is used in three different stages of the USRF framework: to compute the effectiveness estimation; to compute the correlation measure; and to fuse the rankers (through CPRR [76]).

A joint analysis of the parameters k and β was conducted for each one of the datasets considering the custom scenario. The results are presented by Figure 4. To perform this experiment, the Authority Score was used as the effectiveness estimation measure and RBO as the correlation measure, once they presented results comparable or superior to the other measures in [66] and [53], respectively.

[2] `featureselection.asu.edu`



(a) MPEG-7

(b) Flowers

(c) Corel5k

(d) UKBench

(e) Holidays

Figure 4: Impact of parameters K and β on the weighted arithmetic mean considering the MAP of the top-5 selected pairs for each dataset.

Based on the obtained results, we adopted the value of k as (20, 50, 50, 5, 3) for the datasets MPEG-7, Flowers, Corel5k, UKBench and Holidays, respectively. The value of β is analyzed in Section 4.2.3, once it can vary according to the selection scenario. However, observing the obtained results, it can be seen that $\beta = 1$ seems to be satisfactory for the custom scenarios.

4.2.2. Selection Measures

The proposed method is very flexible allowing different effectiveness estimation and rank correlation measures to be used. Table 6 shows the selection results for each one of the datasets considering different measures and the custom scenarios. The reported results consider a weighted arithmetic mean of MAP obtained by the top-5 selected pairs for different combinations of measures.

Notice that, in general, the results present small variations for different measures, what evinces the robustness of our method to the the chosen measures. For most cases, the best results were obtained considering *Reciprocal Density* and *RBO*. Therefore, these two measures were adopted for all of the remaining experiments.

Table 6: Weighted arithmetic mean considering the MAP of the top-5 selected pairs for different combinations of measures (fixed $\beta = 1$).

Measures		**Average MAP for Datasets (%)**				
Effectiveness	**Correlation**	**MPEG-7**	**Flowers**	**Corel5k**	**Ukbench**	**Holidays**
Authority	**Jaccard**	98.95	70.15	**85.93**	96.32	80.14
	Jaccard$_k$	99.33	72.82	**85.93**	97.07	**86.64**
	RBO	99.33	**73.39**	**85.93**	96.97	86.53
	Kendallτ	99.33	73.37	**85.93**	97.07	**86.64**
	Spearman	99.09	70.15	85.60	97.02	**86.64**
Reciprocal	**Jaccard**	99.50	70.44	**85.93**	97.06	84.77
	Jaccard$_k$	**99.62**	70.44	**85.93**	97.15	**86.64**
	RBO	**99.62**	72.52	**85.93**	**97.43**	86.53
	Kendallτ	**99.62**	72.52	**85.93**	97.34	**86.64**
	Spearman	99.60	70.44	85.60	97.34	**86.64**

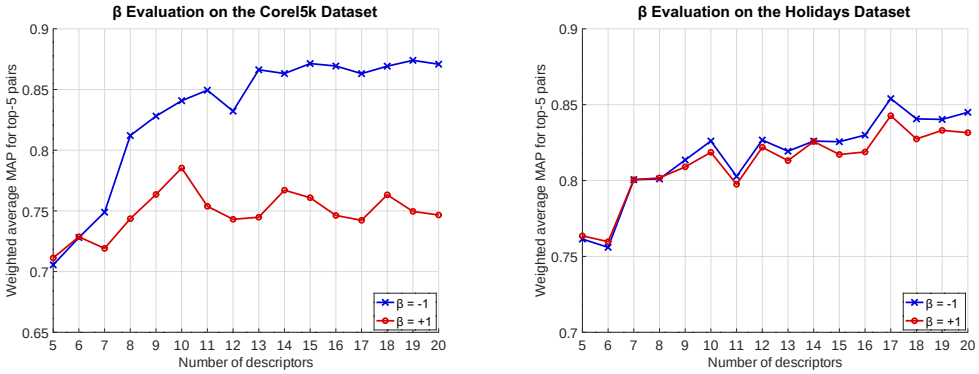
4.2.3. Correlation Relevance

The parameter β adjusts the weight of relevance on the selection measure. In previous experiments (Section 4.2.1), it was seen that $\beta = 1$ seemed to be

satisfactory. However, the value of β is relatively sensible to the number of descriptors offered as input. Therefore, we conducted an experiment which evaluates the effectiveness of our approach for randomly generated scenarios of different sizes. The results are presented in Figure 5 for two datasets. Each of the dots in the graph corresponds to the mean of 20 executions. Once again, for each execution, we consider the weighted average of the top-5 selected pairs.

In general, the results reveal that for scenarios with few descriptors (around 6 or less) $\beta = 1$ is satisfactory. However, as the number of descriptors increase, $\beta = -1$ presents even better results. This is due to the fact that, with a small set of descriptors, the USRF is capable of using the correlation as a way of exploiting the complementarity of the data. While for a large set of descriptors, the correlation is more adequate when selecting the most similar elements, once this strategy filters the outliers.

Therefore, for the remaining experiments, we adopted $\beta = 1$ for the custom scenarios (six descriptors) and $\beta = -1$ for the others.



(a) Corel5k (b) Holidays

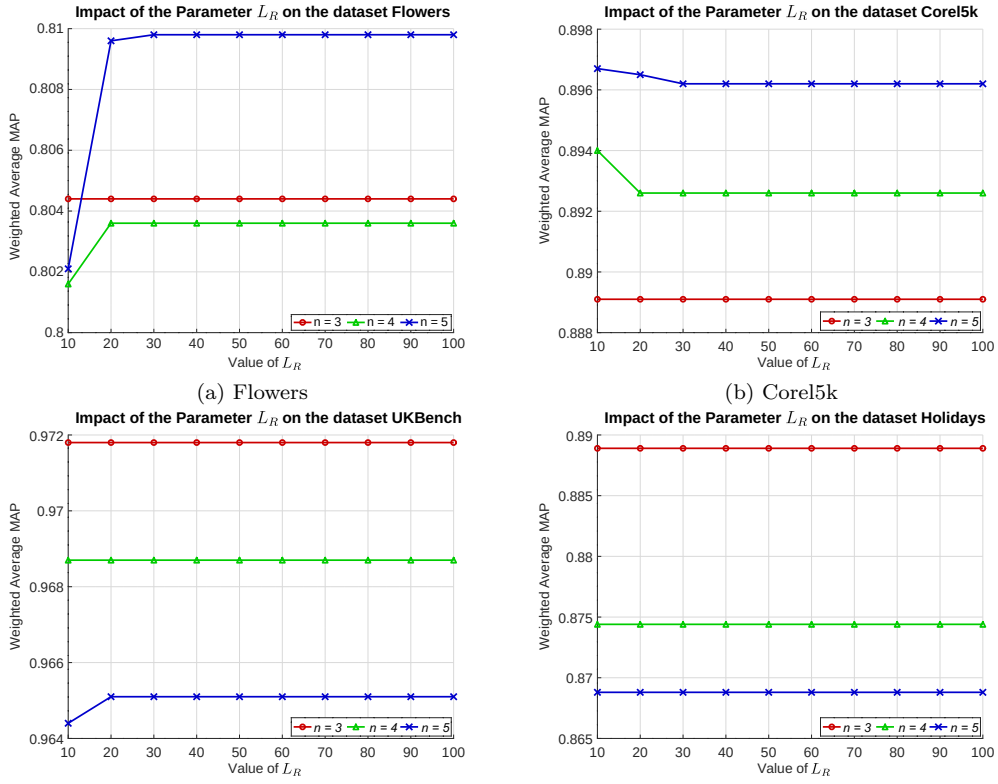
Figure 5: Evaluation of the parameter β considering randomly generated scenarios with different number of descriptors.

4.2.4. Size of the Ranked Combinations Lists

The ranked combinations lists τ_n^R store the combinations sorted by decreasing order of selection score w. As the number of combinations (n) raises, the number of possible combinations exponentially increases. Therefore, it becomes crucial to not consider the whole τ_n^R, but just its first top-L_R positions.

An experiment was performed aiming at evaluating the impact of the parameter L_R on the results. In this analysis, we considered the weighted arithmetic mean of the MAP of the top-5 elements in τ_n^R with n (size of the combination) in the interval $[3, 5]$. For all the datasets, all the descriptors were considered (full scenario).

The results are shown in Figure 6. For a better visualization of the data, the graphs are shown separately due to the scale of the values. As can be seen, there is a very small variation as L_R increases. Therefore, we concluded that the value of L_R can not be very low ($L_R < 30$) to not compromise the selection results. We adopted $L_R = 100$ for all the remaining experiments, once high execution times occur only with values a lot higher than that.



(a) Flowers (b) Corel5k

(c) UKBench (d) Holidays

Figure 6: Impact of the parameter L_R on the selection results.

4.3. Selection Results

While Section 4.3.1 presents analysis for the selection of pairs, Section 4.3.2 reports experiments for combinations of any size. In both cases, different se-

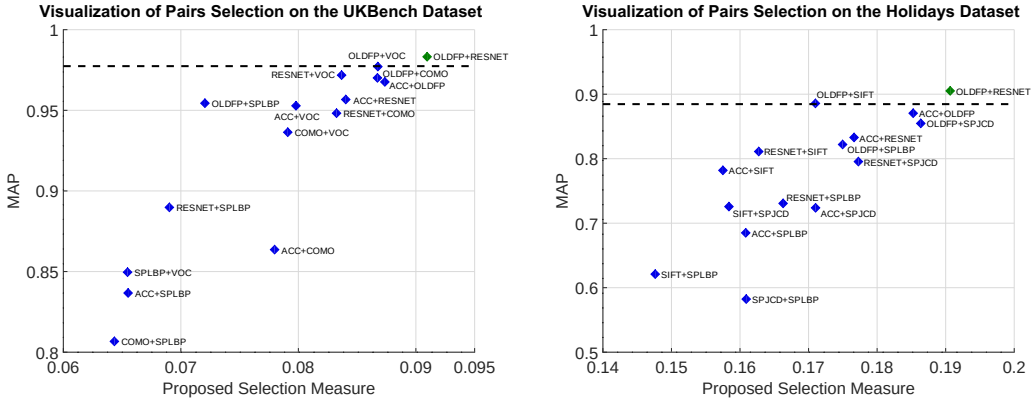
lection scenarios are considered.

4.3.1. Selection of Ranker Pairs

Since the proposed approach is based on the selection of pairs, this section has the objective of primarily evaluate the proposed approach considering only pairs of rankers.

Initially, we report the results considering the custom scenarios. Figure 7 presents the results for the selection of pairs for two different datasets. Each point in the graph corresponds to a different pair of rankers which has the location set according to the value of the selection measure (w_p) computed by the USRF and the MAP obtained for the result of the fused pair.

It is expected that, the higher the value of the selection measure, higher the MAP of the result. The Pearson correlation between the MAP and the selection measure is 0.88 and 0.78 for the graphs (a) and (b), respectively. Such values indicate a strong linear correlation, evidencing the effectiveness of the proposed selection measure w_p.



(a) UKBench (b) Holidays

Figure 7: Distribution of the pairs considering the proposed unsupervised selection measure compared with MAP in the custom scenarios.

For comparative purposes, in the next experiments three hypothetical baselines are considered. The objective is to visualize how the proposed approach compares to each one the following cases:

- **Best Case:** the best selection case is the one that always selects the pair with the best MAP among the pairs available;

- **Average Case:** from all the available pairs, we select the one that is in the median if all of them were sorted by the MAP
- **Worst Case:** the worst selection case is the one that always selects the pair with the worst MAP among the pairs available.

Aiming to facilitate the visualization of the results, the following lines are presented: three dashed lines considering the hypothetical baselines (best, average and worst case), a dashed line considering the best isolated visual feature and, finally, a line for the USRF. The horizontal axis corresponds to the number of selected pairs and the vertical axis indicates the arithmetic mean of the MAP of the selected pairs. This representation evaluates how our approach is compared to each case.

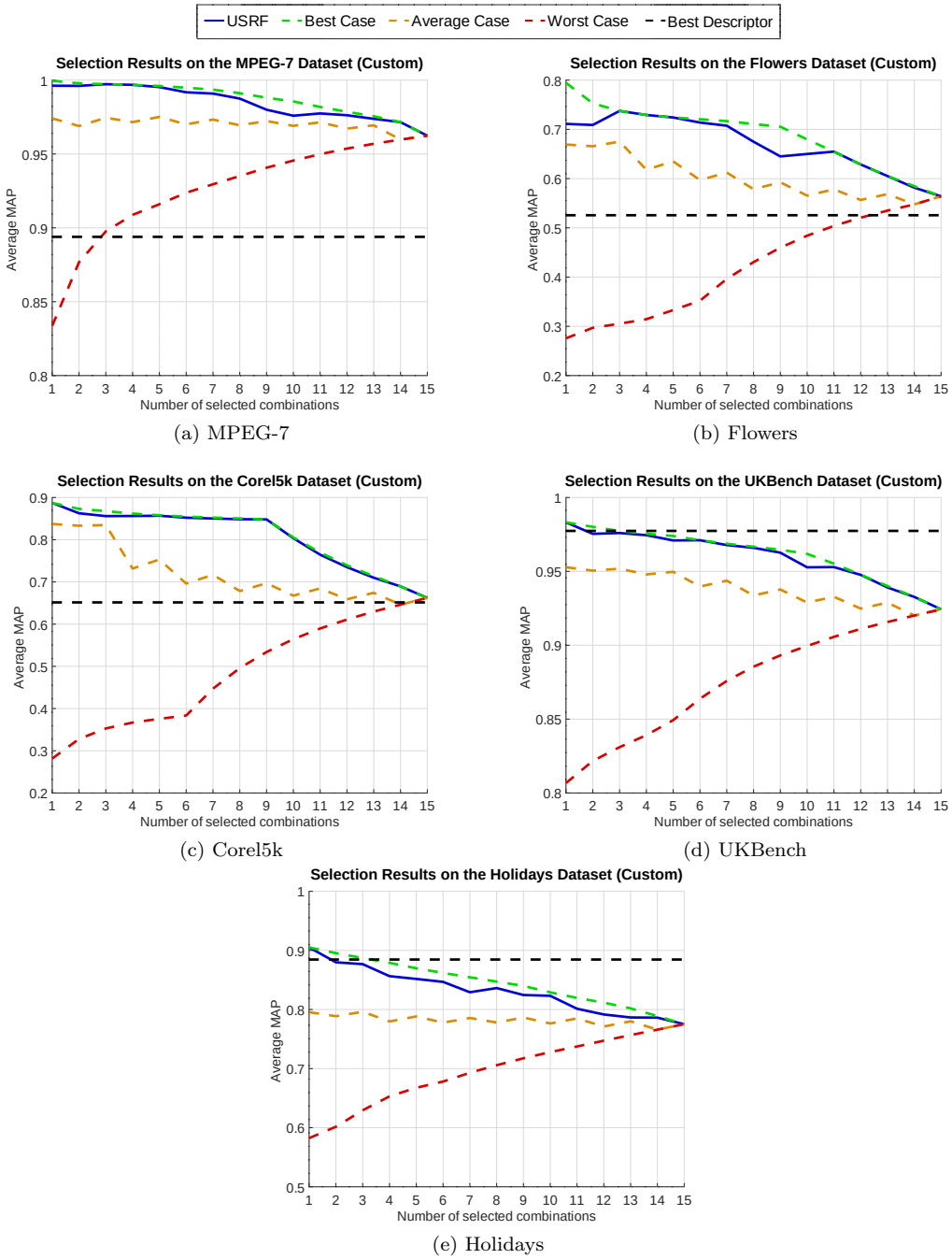
Figure 8 reports the selection results for the Custom Scenario, considering each dataset. It is noticeable that USRF has achieved results comparable to the best case in all the datasets. Notice that the best case is a very strict criteria, once it is always based on the MAP while the USRF is completely unsupervised. In addition, the results obtained by the USRF for a small number of pairs is superior the the best isolated ranker for all datasets.

The same experiment was reproduced considering all the rankers for each dataset, which define our Full Scenario. Figure 9 presents the results for each dataset (except for MPEG-7, where the custom scenario is identical to the full scenario). The scenarios with all descriptors are more challenging, once the variation among the rankers tends to be higher. Once again, the selection results are very similar to the best case on all datasets. Notice that, for the datasets UKBench and Holidays, there are few pairs with results above the best isolated descriptor and, despite of that, our method was capable of selecting them.

Other more specific selection scenarios were also evaluated. Figure 10 presents the evaluation for the datasets Corel5K and UKBench in two scenarios: (i) using only global and local descriptors; (ii) and only deep learning descriptors. The selection results of the graph (a) are located near of the average case, probably due to the high variability between the MAP of the descriptors. Besides that, pairs with results above of the best isolated descriptor were selected among the first positions.

4.3.2. Selection of Rankers Set

The rankers that compose the best selected combination ($\mathfrak{X}^*$) for each dataset are presented in Table 7. The relative gain is computed in relation



(a) MPEG-7

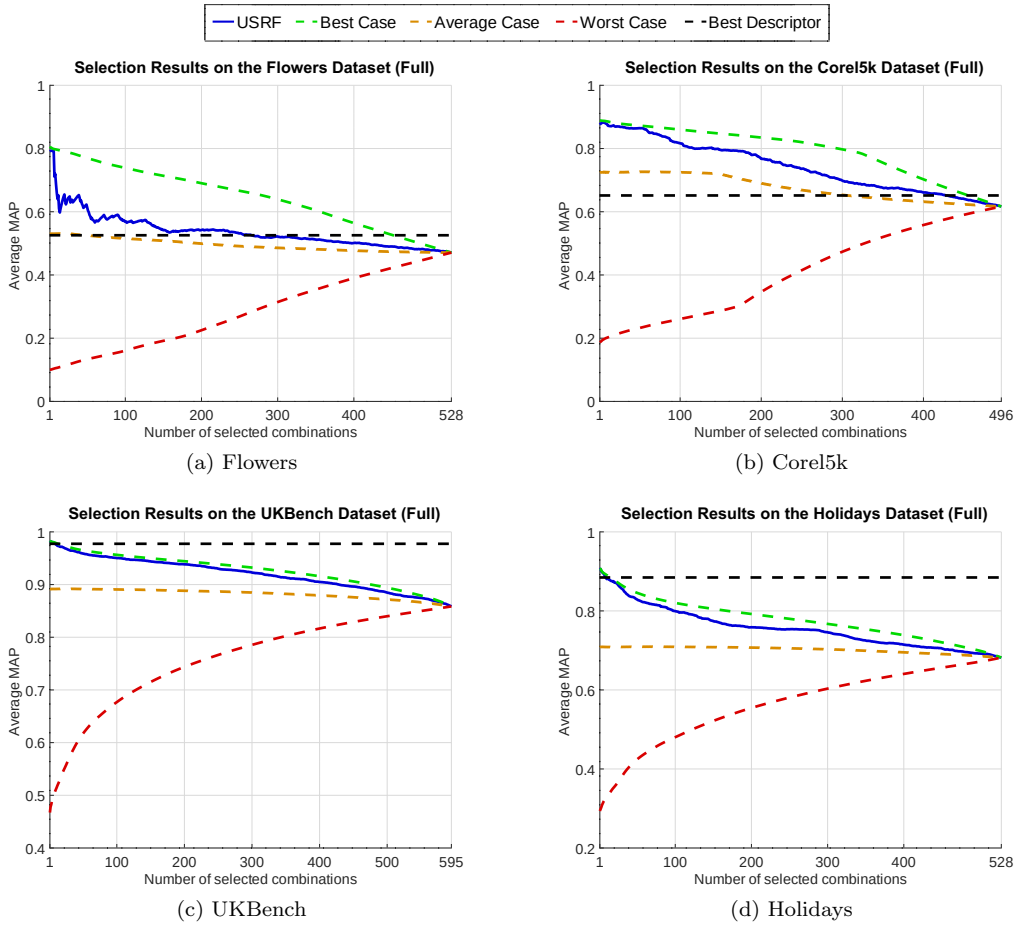
(b) Flowers

(c) Corel5k

(d) UKBench

(e) Holidays

Figure 8: Evaluation of the proposed approach (USRF) for pairs of rankers on the custom scenarios.



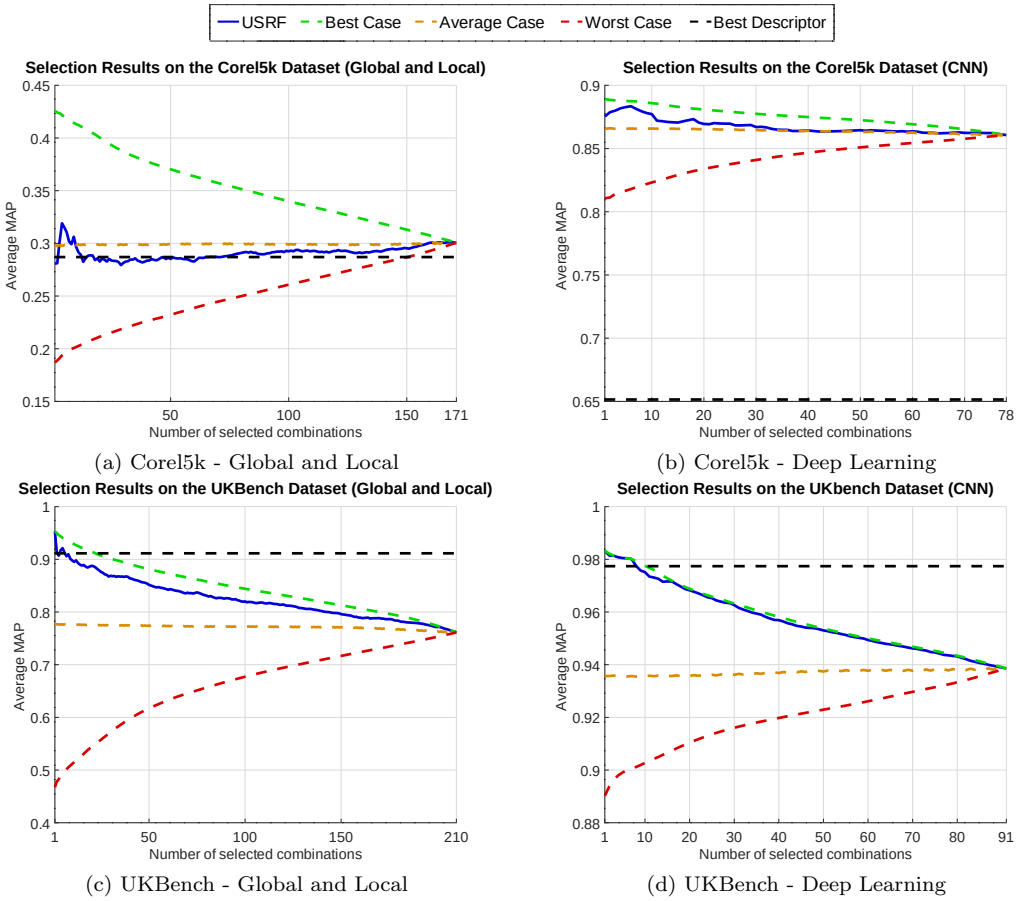
(a) Flowers (b) Corel5k

(c) UKBench (d) Holidays

Figure 9: Evaluation of the proposed approach (USRF) for pairs of rankers on the full scenarios.

to the best isolated descriptor in each case. Notice that the USRF achieved positive gains in all of the presented circumstances.

An experiment was conducted with the purpose of analyzing the impact of the size of the combinations in the USRF. For each dataset, the weighted arithmetic mean of the top-5 combinations was computed (weight 5 for the first position, weight 4 for the second, and so on). The results are reported in Figure 11. For a better visualization of the results, the graphs were splitted due to the difference in the scale of the values. As can be seen, it is not possible to establish a fixed cardinality for $\mathfrak{X}^*$, once the size that offers the best value differs considerably among the datasets. While the datasets Holidays



(a) Corel5k - Global and Local

(b) Corel5k - Deep Learning

(c) UKBench - Global and Local

(d) UKBench - Deep Learning

Figure 10: Evaluation of the proposed approach (USRF) for pairs of rankers on the other scenarios.

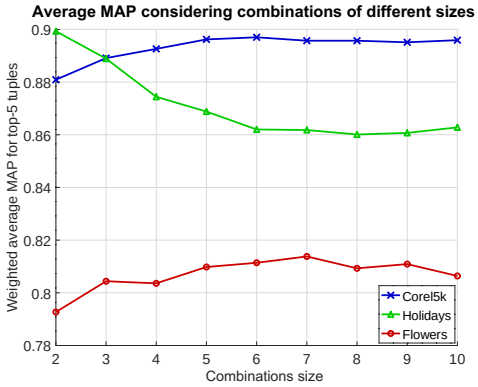
and UKBench are the best values were offered by the combinations of size two (pairs), the others require larger values. Investigating an unsupervised strategy to select the best size for the $\mathfrak{X}^*$ is among of our proposals for future works.

4.4. Comparison with Other Methods

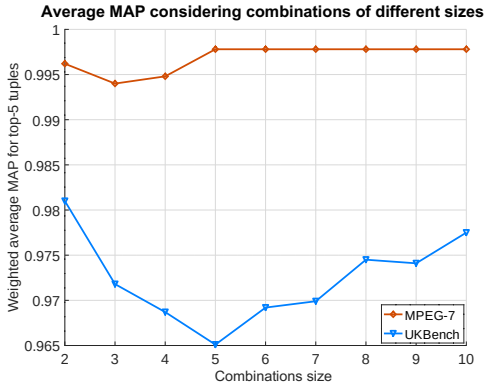
This section presents the comparison of the USRF results with methods of the literature. Section 4.4.1 compares only with techniques based on late fusion and Section 4.4.2 with techniques of feature selection based on early strategies. Finally, Section 4.4.3 confronts our results with the state-of-the-art.

Table 7: MAP of the selected combination for each dataset considering different scenarios.

Scenario	Dataset	Selected Combination	MAP (%)	Relative Gain
Full	**MPEG-7**	AIR + ASC + BAS + CFD + IDSC	99.92	**+11.78%**
	Flowers	AlexNet + BnInception + DPNet + FBResNet + ResNeXt + ResNet + Xception + LBP	81.71	**+55.46%**
	Corel5k	DPNet + FBResNet + InceptionResNet + InceptionV4 + ResNeXt + ResNet + SENet	89.88	**+37.96%**
	UKBench	OLDFP + ResNet	98.32	**+0.59%**
	Holidays	OLDFP + ResNet	90.51	**+2.32%**
Custom	**MPEG-7**	AIR + ASC + BAS + CFD + IDSC	99.92	**+11.78%**
	Flowers	BIC + FBResNet + ResNeXt + SIFT	79.10	**+50.50%**
	Corel5k	DPNet + ResNet + SPACC	90.32	**+38.63%**
	UKBench	ACC + OLDFP + ResNet + VOC	99.02	**+1.31%**
	Holidays	OLDFP + ResNet	90.51	**+2.32%**



(a) Corel5k, Holidays, Flowers



(b) MPEG-7, UKBench

Figure 11: Analysis of the size of the combinations for each of the datasets considering all the rankers (full scenarios).

4.4.1. Late Fusion

For all the baselines, we considered the parameters reported by the authors in their original work. Regarding cases where the dataset was not evaluated in the original paper, we used the same value of k adopted in the USRF, with the intent of making a fair comparison. Aiming to guarantee the executions for all the datasets, the results are presented only for the custom scenarios.

The comparison of the proposed approach with different late fusion baselines is shown by Table 8. It can be seen that our approach presents results better than most of the datasets.

Table 8: USRF compared to late fusion baselines on the custom scenarios.

Method	MAP (%)				
	MPEG-7	**Flowers**	**Corel5k**	**UKBench**	**Holidays**
Best Descriptor	89.39	52.56	65.15	97.74	88.46
Correlation Graph **[63]**	95.79	46.22	58.25	96.06	65.11
Query-Adaptive Fusion **[98]**	92.03	49.56	69.60	96.61	82.59
Graph Fusion **[95]**	99.05	53.60	68.21	**99.24**	85.71
Proposed Method (USRF)	**99.92**	**79.10**	**90.32**	99.02	**90.51**

4.4.2. Late Fusion with Feature Selection

After applying PCA and L2-norm regularization for all the feature vectors, the feature selection methods were employed to select the top-100 most relevant features among the available ones. From the Euclidean distance of the new vectors, we computed the ranked lists for the results and submitted them as input for the CPRR re-ranking algorithm. Besides that, as these methods are very expensive for a large number of features, the results are presented just for the Flowers and Corel5k datasets. A comparison of the proposed approach with feature selection methods, early fusion, are presented in Table 9.

Table 9: USRF compared to feature selection techniques on the custom and full scenarios.

Method	MAP (%)			
	Flowers		**Corel5k**	
	Custom	**Full**	**Custom**	**Full**
Best Descriptor	52.56		65.15	
Laplace **[27]**	70.84	61.28	86.09	78.40
SPEC [96]	71.46	49.67	72.82	63.99
MCFS [10]	75.81	55.95	86.38	84.74
UDFS [93]	69.70	63.97	80.77	78.20
NDFS [41]	71.68	65.41	86.09	87.17
USRF	**79.10**	**81.71**	**90.32**	**89.88**

4.4.3. State-of-the-art

Finally, USRF is evaluated in comparison to the main state-of-the-art methods, among them post-processing approaches and image retrieval methods. The comparisons are presented for the datasets: MPEG-7, Holidays, and UKBench. These datasets are commonly used as benchmark for content-based image retrieval. **Most of the state-of-the-art methods report results considering a pre-selected set of descriptors as input, while**

Table 10: Comparison with various post-processing methods on the MPEG-7 [37] dataset (Bull's eye score - Recall@40).

Shape Descriptors		**Bull's eye score**
Beam Angle Statistics (BAS) [1]	-	75.21%
Contour Feat. Descriptor (CFD) [60]		84.43%
Inner Dist. Shape Context (IDSC) [42]		85.40%
Aspect Shape Context (ASC) [43]		88.39%
Articulation-Invariant Rep. (AIR) [25]		93.67%
Post-Processing Methods	**Descriptors**	**Bull's eye score**
Contextual Dissimilarity Measure [32]	IDSC [42]	88.30%
Graph Transduction [90]		91.00%
Self-Smoothing Operator [33]		92.77%
Local Constr. Diff. Process [91]		93.32%
Mutual kNN Graph [35]		93.40%
SCA [4]		93.44%
Smooth Neighborhood [6]		93.52%
Reciprocal kNN Graph CCs [64]		93.62%
Graph Fusion [95]	CFD [60]	89.76%
Index-Based Re-Ranking [58]		92.85%
RL-Sim [61]		94.27%
Correlation Graph [63]		94.84%
Reciprocal kNN Graph CCs [64]		96.51%
Generic Diffusion Process [21]	ASC [43]	93.95%
Index-Based Re-Ranking [58]		94.09%
Correlation Graph [63]		95.50%
Local Constr. Diff. Process [91]		95.96%
Smooth Neighborhood [6]		95.98%
Reciprocal kNN Graph CCs [64]		96.04%
Tensor Product Graph [92]		96.47%
Graph Fusion [95]	AIR [25]	98.76%
Index-Based Re-Ranking [58]		99.93%
RL-Sim [61]		99.94%
Tensor Product Graph [92]		99.99%
Generic Diffusion Process [21]		100%
Neighbor Set Similarity [8]		100%
Reciprocal kNN Graph CCs [64]		100%
Proposed Approach (USRF)	Selected Comb.	**100%**

our approach is responsible by doing the selection in a completely unsupervised fashion. Therefore, we highlight that the comparison with such state-of-the-art results is a very strict criterion.

Table 10 presents the results for the MPEG-7 dataset considering the Recall@40, also known as bull's eye score. As can be seen, the USRF is capable of selecting a combination with maximum value of **100%**.

The results obtained on the Holidays dataset are shown in Table 11. Several state-of-the-art methods are included in the comparison. Notice that USRF achieved a MAP of **90.51%** in both scenarios, being comparable to

the best retrieval results. It can be seen that the larger number of rankers available in the full scenario did not compromise the selection, obtaining results comparable or superior to the state-of-the-art in both cases.

Table 11: Comparison with the state-of-the-art in the Holidays dataset (MAP).

MAP for the state-of-the-art methods				
Jégou *et al.* [31]	Tolias *et al.* [75]	Paulin *et al.* [57]	Qin *et al.* [68]	Zheng *et al.* [99]
75.07%	82.20%	82.90%	84.40%	85.20%

Sun *et al.* [72]	Zheng *et al.* [97]	Pedronette et al. [64]	Li *et al.* [40]	Liu et al. [46]
85.50%	85.80%	86.19%	89.20%	**90.89%**

Proposed Method (USRF)	
Custom Scenario	90.51%
Full Scenario	90.51%

Similarly, Table 12 presents a comparison on the UKBench dataset. Our proposed method achieved a N-S Score, which is similar to P@4, of 3.90 in the custom scenario and **3.94** in the full scenario. The full scenario is more challenging due to the large number of rankers available. Besides that, the USRF achieved a result even more significant in the full scenario, which reveals the accuracy of the proposed selection algorithm.

Table 12: Comparison with the state-of-the-art in the UKBench dataset (N-S Score).

***N-S scores* for the state-of-the-art methods**					
Zheng et al. [100]	Wang et al. [80]	Sun et al. [72]	Paulin *et al.* [57]	Zhang et al. [95]	Zheng et al. [98]
3.57	3.68	3.76	3.76	3.83	3.84

Bai et al. [4]	Xie et al. [84]	Liu et al. [46]	Pedronette et al. [64]	Bai et al. [7]	Bai et al. [5]
3.86	3.89	3.92	3.93	3.93	**3.94**

Proposed Method (USRF)	
Custom Scenario	3.90
Full Scenario	**3.94**

4.5. Visual Results

With the purpose of offering a qualitative visualization of the results, the Figure 12 illustrates two example of ranked lists that compare the individual

results for each ranker (descriptor) and the obtained by the combination selected by the USRF. The query images are presented in green borders and the incorrect results in red borders.

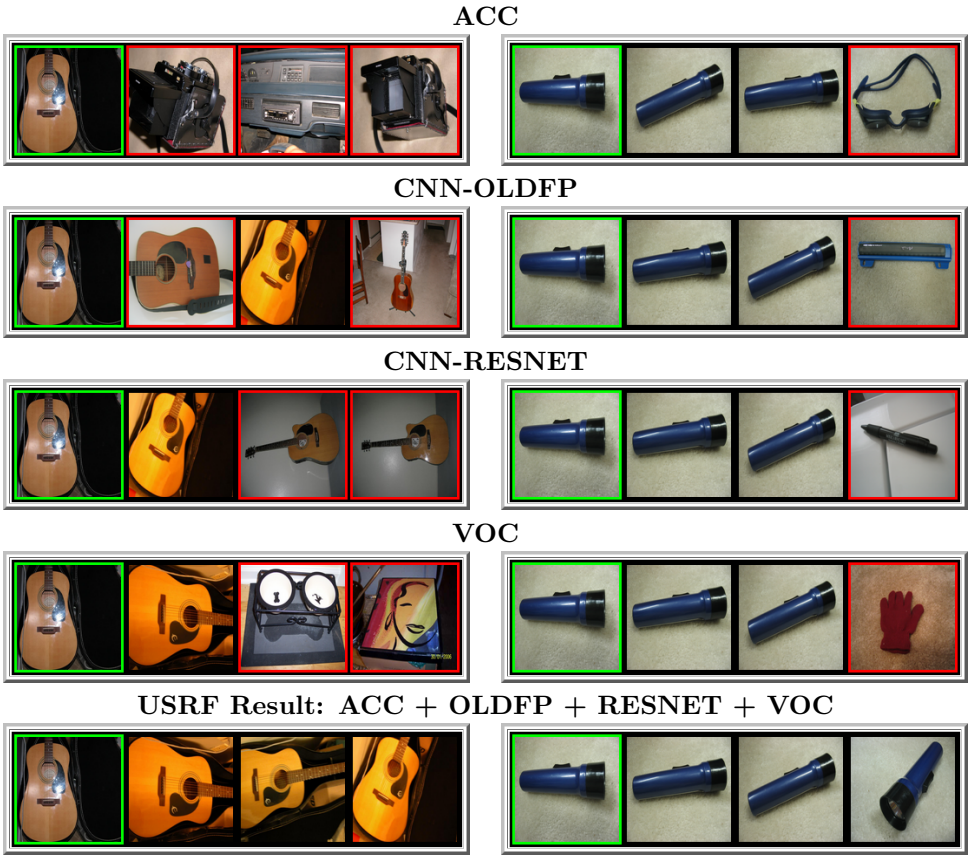


Figure 12: Two visual results showing the impact of our proposed selection and rank fusion approach on UKBench dataset.

5. Conclusions

Selecting and fusing descriptors is of crucial relevance in image retrieval tasks, especially in unsupervised scenarios. In this paper we have presented an approach to select and combine descriptors (or rankers) in a completely unsupervised way. Based on rank correlation and effectiveness estimation measures, the proposed framework is flexible, and allow the use of different

measures. A broad experimental evaluation was conducted, involving various experiments, different retrieval tasks and several datasets and image features. Experimental results demonstrated the potential of our approach in selecting high-effective combinations. As future work, we intend to investigate the unsupervised determination of the best size of combinations to be used by the proposed USRF approach.

Acknowledgments

The authors are grateful to São Paulo Research Foundation - FAPESP (grants #2017/02091-4, #2018/15597-6, and #2017/25908-6), Brazilian National Council for Scientific and Technological Development - CNPq (grant #308194/2017-9) and Microsoft Research.

References

[1] N. Arica, F. T. Y. Vural, BAS: a perceptual shape descriptor based on the beam angle statistics, Pattern Recognition Letters 24 (9-10) (2003) 1627–1639.

[2] K. S. Arun, V. K. Govindan, S. D. M. Kumar, On integrating re-ranking and rank list fusion techniques for image retrieval, International Journal of Data Science and Analytics 4 (1) (2017) 53–81.

[3] P. K. Atrey, M. A. Hossain, A. El Saddik, M. S. Kankanhalli, Multimodal fusion for multimedia analysis: a survey, Multimedia Systems 16 (6) (2010) 345–379, ISSN 1432-1882.

[4] S. Bai, X. Bai, Sparse Contextual Activation for Efficient Visual Re-Ranking, IEEE Trans. on Image Processing (TIP) 25 (3) (2016) 1056–1069.

[5] S. Bai, X. Bai, Q. Tian, L. J. Latecki, Regularized Diffusion Process for Visual Retrieval, in: Conf. on Artificial Intelligence (AAAI), 3967–3973, 2017.

[6] S. Bai, S. Sun, X. Bai, Z. Zhang, Q. Tian, Smooth Neighborhood Structure Mining on Multiple Affinity Graphs with Applications to Context-Sensitive Similarity, in: European Conference on Computer Vision (ECCV), 592–608, 2016.

[7] S. Bai, Z. Zhou, J. Wang, X. Bai, L. J. Latecki, Q. Tian, Ensemble Diffusion for Retrieval, in: 2017 IEEE International Conference on Computer Vision (ICCV), 774–783, 2017.

[8] X. Bai, S. Bai, X. Wang, Beyond diffusion process: Neighbor set similarity for fast re-ranking, Information Sciences 325 (2015) 342 – 354.

[9] S. S. Bucak, R. Jin, A. K. Jain, Multiple Kernel Learning for Visual Object Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (7) (2014) 1354–1369.

[10] D. Cai, C. Zhang, X. He, Unsupervised Feature Selection for Multi-cluster Data, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, ISBN 978-1-4503-0055-1, 333–342, 2010.

[11] S. A. Chatzichristofis, Y. S. Boutalis, CEDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval, in: Proceedings of the 6th international conference on Computer vision systems, ICVS'08, 312–322, 2008.

[12] S. A. Chatzichristofis, Y. S. Boutalis, FCTH: Fuzzy Color and Texture Histogram A Low Level Feature for Accurate Image Retrieval, in: Int. Workshop on Image Analysis for Multimedia Interactive Services, 191–196, 2008.

[13] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, J. Feng, Dual Path Networks, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 4467–4475, 2017.

[14] F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1800–1807, 2017.

[15] L. Cieplinski, MPEG-7 Color Descriptors and Their Applications, in: W. Skarbek (Ed.), Computer Analysis of Images and Patterns, Springer Berlin Heidelberg, Berlin, Heidelberg, 11–20, 2001.

[16] R. da S. Torres, A. X. Falcão, Contour Salience Descriptors for Effective Image Retrieval and Analysis, Image and Vision Computing 25 (1) (2007) 3–13.

[17] R. da S. Torres, A. X. Falcão, M. A. Gonçalves, J. P. Papa, B. Zhang, W. Fan, E. A. Fox, A genetic programming framework for content-based image retrieval, vol. 42, ISSN 0031-3203, 283 – 292, learning Semantics from Multimedia Content, 2009.

[18] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, 886–893 vol. 1, 2005.

[19] R. Datta, D. Joshi, J. Li, J. Z. Wang, Image retrieval: Ideas, influences, and trends of the new age, ACM Computing Surveys 40 (2) (2008) 5:1–5:60.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09, 2009.

[21] M. Donoser, H. Bischof, Diffusion Processes for Retrieval Revisited, in: Conf. on Computer Vision and Pattern Recognition (CVPR), 1320–1327, 2013.

[22] C. Dwork, R. Kumar, M. Naor, D. Sivakumar, Rank Aggregation Methods for the Web, in: Proceedings of the 10th International Conference on World Wide Web, WWW '01, ISBN 1-58113-348-0, 613–622, 2001.

[23] R. Fagin, R. Kumar, D. Sivakumar, Comparing top k lists, in: ACM-SIAM Symposium on Discrete algorithms (SODA'03), ISBN 0-89871-538-5, 28–36, 2003.

[24] C. D. Ferreira, J. A. dos Santos, R. da S. Torres, M. A. Gonçalves, R. C. Rezende, W. Fan, Relevance feedback based on genetic programming for image retrieval, Pattern Recogninion Letters 32 (1) (2011) 27–37, ISSN 0167-8655.

[25] R. Gopalan, P. Turaga, R. Chellappa, Articulation-invariant representation of non-planar shapes, in: 11th European Conference on Computer Vision (ECCV'2010), vol. 3, 286–299, 2010.

[26] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778, 2016.

[27] X. He, D. Cai, P. Niyogi, Laplacian Score for Feature Selection, in: Proceedings of the 18th International Conference on Neural Information Processing Systems, NIPS'05, 507–514, 2005.

[28] J. Hu, L. Shen, G. Sun, Squeeze-and-Excitation Networks, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[29] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, R. Zabih, Image Indexing Using Color Correlograms, in: CVPR'97, ISBN 0-8186-7822-4, 762–768, 1997.

[30] S. Ioffe, C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in: Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, 448–456, 2015.

[31] H. Jegou, M. Douze, C. Schmid, Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search, in: European Conference on Computer Vision, ECCV '08, 304–317, 2008.

[32] H. Jegou, C. Schmid, H. Harzallah, J. Verbeek, Accurate Image Search Using the Contextual Dissimilarity Measure, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (1) (2010) 2–11.

[33] J. Jiang, B. Wang, Z. Tu, Unsupervised metric learning by Self-Smoothing Operator, in: Int. Conference on Computer Vision (ICCV), 794–801, 2011.

[34] M. L. Kherfi, D. Ziou, Relevance feedback for CBIR: a new approach based on probabilistic feature weighting with positive and negative examples, IEEE Transactions on Image Processing 15 (4) (2006) 1017–1030.

[35] P. Kontschieder, M. Donoser, H. Bischof, Beyond Pairwise Shape Similarity Analysis, in: Asian Conf. on Computer Vision (ACCV'2009), 655–666, 2009.

[36] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12, 1097–1105, 2012.

[37] L. J. Latecki, R. Lakmper, U. Eckhardt, Shape Descriptors for Non-rigid Shapes with a Single Closed Contour, in: CVPR'2000, 424–429, 2000.

[38] M. S. Lew, N. Sebe, C. Djeraba, R. Jain, Content-based Multimedia Information Retrieval: State of the Art and Challenges, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 2 (1) (2006) 1–19.

[39] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, H. Liu, Feature Selection: A Data Perspective, arXiv preprint arXiv:1601.07996 .

[40] X. Li, M. Larson, A. Hanjalic, Pairwise geometric matching for large-scale object retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2015), 5153–5161, 2015.

[41] Z. Li, Y. Yang, J. Liu, X. Zhou, H. Lu, Unsupervised Feature Selection Using Nonnegative Spectral Analysis, in: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.

[42] H. Ling, D. W. Jacobs, Shape Classification Using the Inner-Distance, IEEE Trans. on Pattern Analysis and Machine Intell. 29 (2) (2007) 286–299, ISSN 0162-8828.

[43] H. Ling, X. Yang, L. J. Latecki, Balancing Deformability and Discriminability for Shape Matching, in: ECCV'2010, vol. 3, 411–424, 2010.

[44] G.-H. Liu, J.-Y. Yang, Content-based image retrieval using color difference histogram, Pattern Recognition 46 (1) (2013) 188 – 198.

[45] S. Liu, W. Deng, Very deep convolutional neural network based image classification using small training sample size, in: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), 730–734, 2015.

[46] Z. Liu, S. Wang, L. Zheng, Q. Tian, Robust ImageGraph: Rank-Level Feature Fusion for Image Search, IEEE Transactions on Image Processing 26 (7) (2017) 3128–3141.

[47] D. Lowe, Object recognition from local scale-invariant features, in: IEEE International Conference on Computer Vision (ICCV), 1150–1157, 1999.

[48] M. Lux, Content Based Image Retrieval with LIRe, in: Proceedings of the 19th ACM International Conference on Multimedia, MM '11, 2011.

[49] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, A. Yamada, Color and texture descriptors, IEEE Transactions on Circuits and Systems for Video Technology 11 (6) (2001) 703–715.

[50] M.-E. Nilsback, A. Zisserman, A Visual Vocabulary for Flower Classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 1447–1454, 2006.

[51] D. Nistér, H. Stewénius, Scalable Recognition with a Vocabulary Tree, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2006), vol. 2, 2161–2168, 2006.

[52] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7) (2002) 971–987, ISSN 0162-8828.

[53] C. Y. Okada, D. C. G. Pedronette, R. da S. Torres, Unsupervised Distance Learning by Rank Correlation Measures for Image Retrieval, in: ACM International Conference on Multimedia Retrieval (ICMR'2015), 2015.

[54] A. Oliva, A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, IJCV 42 (3) (2001) 145–175.

[55] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, 1999.

[56] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch, in: NIPS-W, 2017.

[57] M. Paulin, J. Mairal, M. Douze, Z. Harchaoui, F. Perronnin, C. Schmid, Convolutional Patch Representations for Image Retrieval: An Unsupervised Approach, Int. Journal of Computer Vision .

[58] D. C. G. Pedronette, J. Almeida, R. da S. Torres, A Scalable Re-Ranking Method for Content-Based Image Retrieval, Information Sciences 265 (1) (2014) 91–104.

[59] D. C. G. Pedronette, R. d. S. Torres, Unsupervised Effectiveness Estimation for Image Retrieval Using Reciprocal Rank Information, in: 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images, 321–328, 2015.

[60] D. C. G. Pedronette, R. da S. Torres, Shape Retrieval using Contour Features and Distance Optmization, in: VISAPP'2010, vol. 1, 197 – 202, 2010.

[61] D. C. G. Pedronette, R. da S. Torres, Image Re-Ranking and Rank Aggregation based on Similarity of Ranked Lists, Pattern Recognition 46 (8) (2013) 2350–2360.

[62] D. C. G. Pedronette, R. da S. Torres, Unsupervised Effectiveness Estimation for Image Retrieval using Reciprocal Rank Information, in: Conference on Graphics, Patterns and Images (SIBGRAPI'2015), 2015.

[63] D. C. G. Pedronette, R. da S. Torres, A correlation graph approach for unsupervised manifold learning in image retrieval tasks, Neurocomputing 208 (Sup C) (2016) 66 – 79.

[64] D. C. G. Pedronette, F. M. F. Gonçalves, I. R. Guilherme, Unsupervised manifold learning through reciprocal kNN graph and Connected Components for image retrieval tasks, Pattern Recognition 75 (2018) 161 – 174.

[65] D. C. G. Pedronette, O. A. Penatti, R. da S. Torres, Unsupervised manifold learning using Reciprocal kNN Graphs in image re-ranking and rank aggregation tasks, Image and Vision Computing 32 (2) (2014) 120 – 130.

[66] D. C. G. Pedronette, O. A. Penatti, R. da S. Torres, Unsupervised manifold learning using Reciprocal kNN Graphs in image re-ranking

and rank aggregation tasks, Image and Vision Computing 32 (2) (2014) 120 – 130.

[67] L. Piras, G. Giacinto, Information fusion in content based image retrieval: A comprehensive overview, Information Fusion 37 (Supplement C) (2017) 50 – 60.

[68] D. Qin, C. Wengert, L. V. Gool, Query Adaptive Similarity for Large Scale Object Retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2013), 1610–1617, 2013.

[69] K. Reddy Mopuri, R. Venkatesh Babu, Object Level Deep Feature Pooling for Compact Image Representation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2015.

[70] C. G. M. Snoek, M. Worring, A. W. M. Smeulders, Early Versus Late Fusion in Semantic Video Analysis, in: Proceedings of the 13th Annual ACM International Conference on Multimedia, MULTIMEDIA '05, ISBN 1-59593-044-2, 399–402, 2005.

[71] R. O. Stehling, M. A. Nascimento, A. X. Falcão, A compact and efficient image retrieval approach based on border/interior pixel classification, in: CIKM'2002, ISBN 1-58113-492-4, 102–109, 2002.

[72] S. Sun, Y. Li, W. Zhou, Q. Tian, H. Li, Local residual similarity for image re-ranking, Information Sciences 417 (Sup. C) (2017) 143 – 153.

[73] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, 2017.

[74] X. Tian, Y. Lu, L. Yang, Query Difficulty Prediction for Web Image Search, Multimedia, IEEE Transactions on 14 (4) (2012) 951–962.

[75] G. Tolias, Y. Avrithis, H. Jégou, To Aggregate or Not to aggregate: Selective Match Kernels for Image Search, in: IEEE International Conference on Computer Vision (ICCV'2013), 1401–1408, 2013.

[76] L. P. Valem, D. C. G. Pedronette, Unsupervised similarity learning through Cartesian product of ranking references, Pattern Recognition Letters On-line, To appear.

[77] L. P. Valem, D. C. G. a. Pedronette, Selection and Combination of Unsupervised Learning Methods for Image Retrieval, in: Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing, CBMI '17, ISBN 978-1-4503-5333-5, 27:1–27:6, 2017.

[78] K. E. A. van de Sande, T. Gevers, C. G. M. Snoek, Evaluating Color Descriptors for Object and Scene Recognition, PAMI 32 (9) (2010) 1582–1596.

[79] S. A. Vassou, N. Anagnostopoulos, A. Amanatiadis, K. Christodoulou, S. A. Chatzichristofis, CoMo: A Compact Composite Moment-Based Descriptor for Image Retrieval, in: Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing, CBMI '17, ISBN 978-1-4503-5333-5, 30:1–30:5, 2017.

[80] B. Wang, J. Jiang, WeiWang, Z.-H. Zhou, Z. Tu, Unsupervised Metric Fusion by Cross Diffusion, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2012), 3013 –3020, 2012.

[81] X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, T. Han, Contextual weighting for vocabulary tree based image retrieval, in: IEEE International Conference on Computer Vision (ICCV'2011), 209–216, 2011.

[82] W. Webber, A. Moffat, J. Zobel, A similarity measure for indefinite rankings, ACM Transactions on Information Systems 28 (4) (2010) 20:1–20:38, ISSN 1046-8188.

[83] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, A. Torralba, SUN database: Large-scale scene recognition from abbey to zoo, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 3485–3492, 2010.

[84] L. Xie, R. Hong, B. Zhang, Q. Tian, Image Classification and Retrieval Are ONE, in: ACM Int. Conf. on Multimedia Retrieval (ICMR), 3–10, 2015.

[85] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated Residual Transformations for Deep Neural Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5987–5995, 2017.

[86] X. Xing, Y. Zhang, M. Han, Query Difficulty Prediction for Contextual Image Retrieval, in: Advances in Information Retrieval, vol. 5993 of *Lecture Notes in Computer Science*, 581–585, 2010.

[87] C. Yan, L. Li, C. Zhang, B. Liu, Y. Zhang, Q. Dai, Cross-modality Bridging and Knowledge Transferring for Image Understanding, IEEE Transactions on Multimedia (2019) 1–1.

[88] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, Q. Dai, STAT: Spatial-Temporal Attention Mechanism for Video Captioning, IEEE Transactions on Multimedia (2019) 1–1.

[89] C. Yan, H. Xie, J. Chen, Z. Zha, X. Hao, Y. Zhang, Q. Dai, A Fast Uyghur Text Detector for Complex Background Images, IEEE Transactions on Multimedia 20 (12) (2018) 3389–3398.

[90] X. Yang, X. Bai, L. J. Latecki, Z. Tu, Improving Shape Retrieval by Learning Graph Transduction, in: European Conference on Computer Vision (ECCV'2008), vol. 4, 788–801, 2008.

[91] X. Yang, S. Koknar-Tezel, L. J. Latecki, Locally constrained diffusion process on locally densified distance spaces with applications to shape retrieval., in: CVPR'2009, 357–364, 2009.

[92] X. Yang, L. Prasad, L. Latecki, Affinity Learning with Diffusion on Tensor Product Graph, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (1) (2013) 28–38.

[93] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, X. Zhou, L2,1-norm Regularized Discriminative Feature Selection for Unsupervised Learning, in: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, IJCAI'11, 1589–1594, 2011.

[94] K. Zagoris, S. Chatzichristofis, N. Papamarkos, Y. Boutalis, Automatic Image Annotation and Retrieval Using the Joint Composite Descriptor, in: 14th Panhellenic Conference on Informatics (PCI), 143–147, 2010.

[95] S. Zhang, M. Yang, T. Cour, K. Yu, D. Metaxas, Query Specific Rank Fusion for Image Retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (4) (2015) 803–815.

[96] Z. Zhao, H. Liu, Spectral Feature Selection for Supervised and Unsupervised Learning, in: Proceedings of the 24th International Conference on Machine Learning, ICML '07, ISBN 978-1-59593-793-3, 1151–1157, 2007.

[97] L. Zheng, S. Wang, Z. Liu, Q. Tian, Packing and Padding: Coupled Multi-index for Accurate Image Retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR'2014), 1947–1954, 2014.

[98] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, Q. Tian, Query-Adaptive Late Fusion for Image Search and Person Re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[99] L. Zheng, S. Wang, Q. Tian, Coupled Binary Embedding for Large-Scale Image Retrieval, IEEE Transactions on Image Processing (TIP) 23 (8) (2014) 3368–3380.

[100] L. Zheng, S. Wang, Q. Tian, Lp-Norm IDF for Scalable Image Retrieval, IEEE Trans. on Image Processing 23 (8) (2014) 3604–3617.

[101] L. Zheng, Y. Yang, Q. Tian, SIFT Meets CNN: A Decade Survey of Instance Retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (5) (2018) 1224–1244.

[102] W. Zhou, H. Li, Q. Tian, Recent Advance in Content-based Image Retrieval: A Literature Survey, CoRR abs/1706.06064.

[103] B. Zoph, V. Vasudevan, J. Shlens, Q. V. Le, Learning Transferable Architectures for Scalable Image Recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), 2018.