Clustering Effectiveness Analyses for Visual Diversity

Iago Breno Alves do Carmo Araujo, *Author,* Rodrigo Tripodi Calumby, *Advisor*

***Abstract*—Image retrieval engines rely on similarity-based processing for finding and retrieving relevant images for a given user-defined query. This relevance-oriented approach, although effective, sometimes does not completely satisfies the possible multiple information needs related to, e.g., ambiguous or under-specified queries or visual summarization tasks. As an alternative, result diversity has been promoted to comprise multiple search aspects. In this context, although many clustering approaches have been applied as part of diversification methods, there is no comprehensive experimental analysis regarding the extent of features and algorithms effectiveness for image clustering and how close they are to human labeling. This paper proposes an analysis based on clustering approaches and discusses experimental results for assessing several visual and textual features and algorithms. Our experimental analysis revealed the difficulty and poor effectiveness of the evaluated methods when automatic clustering-based image groups are compared to the partitions manually created by human assessors. Our findings suggest that the effectiveness limitations of the algorithms and features are a consequence of the semantic gap problem regarding the high level human concepts considered for image clustering and the actual capability of automatic methods on performing such task.**

***Index Terms*—Clustering, Visual Diversity, Content-based Image Retrieval.**

I. Introduction

LARGE collections of images are generated fostered by the technological advances in image acquisition and data storage. Thus, the need to handle these collections in an efficient and effective manner has arisen. In this context, content-based image retrieval systems are introduced as an approach to find the images according to their similarities in relation to a user-provided query example. This approach is centered on the encoding of image visual properties such as color, texture, and shape. Thus, from these properties, it is possible to compute the similarity between the images.

A typical solution for content-based image retrieval requires the application of one or more image descriptors, which can vary according to the application domain. The choice of appropriate image descriptors requires conducting a series of experiments in order to evaluate them in terms of effectiveness for a given image dataset. This evaluation is a complex activity and covers the use of adequate effectiveness assessment measures [1].

As described in [1], content-based image retrieval systems have been used in various applications. An area that extensively uses such systems is medicine. The medical applications can take advantage of theses systems for teaching/learning, researching, and diagnosis. For example, in the case of diagnosis, general practitioners often use the similarity of clinical cases in the decision taking process. In this sense, visual features can be used to retrieve similar information for a certain clinical case. Another example of content-based retrieval application is Biodiversity Information Systems. These systems are based on the work that biologists perform for gathering many biodiversity artifacts, including images of living species. Hence, such systems support researchers in the comprehension and understanding of species and their habitats by the information retrieval according to examples.

Finally, the advances of content-based image retrieval systems led to increasing challenges and difficulties related to the conception and improvement of such systems. The retrieval engine itself has to deal with issues such as the relationship between the high-level perception from the user and low-level visual features extracted from digital objects. In essence, given the images wealth and the subjectivity of human perception, the semantic gap becomes one of the greatest challenges in the retrieval process [2], [3]. Hence, the design of system requires the selection of image descriptors suitable to the application domain, the adequate representation of feature vectors, interaction mechanisms, among others aspects.

Furthermore, the development of image retrieval systems to tackle only relevance-based search has been shown as insufficient to satisfy the myriad of user information needs. According to [4], a query provided by a user is frequently ambiguous in some extent. Moreover, the retrieval results can be redundant. These two aspects, ambiguity and redundancy, have motivated the study of the diversity concept in information retrieval. Therefore, diversifying the retrieval results means to expand the concepts related to a query. Hence, it is expected to improve the coverage and novelty of the retrieval results, making it less prone to wrong guessing regarding the possible information needs. For instance, given a search scenario, a user may provide a query using the Portuguese keyword “minas”, with the intent to catalog mines from a particular region according to the type of deposited mineral. This poor and ambiguous query may not allow the system to satisfy the user looking for different aspects. The system can correctly return relevant images of mineral mines, but very redundant or even near-duplicates. Beyond it, the system could answer to the wrong search intent, exhibiting, e.g., only images of explosive devices related to landmines or even only sights from the Brazilian state of Minas Gerais. Hence, in order to improve the information gain and maximize the coverage of the possible query intents, it is useful to promote the so called visual diversity.

Diversity may be defined as a set property related to the conceptual difference between the items in that given set [5]. It

can be applied to attend some needs, including: (i) the absence of an specific item that satisfies the user; (ii) the user intent of obtaining different concepts from a query; (iii) the user desire to select over a diverse collection of items; and (iv) when many information items need to be combined to provide an adequate result [5]. Moreover, search engines must be flexible to handle ambiguous, poor or complex queries that demand diverse results to be properly answered [6].

In this context, some studies like [7] have shown that relevant information is also associated with diversity and its promotion has allowed proper results with regard to user satisfaction and search experience optimization. Nevertheless, the diversification presents as major drawbackm, which is the possibility of mistakenly promoting irrelevant items to the top of the search results. Therefore, finding a balance between relevance and diversity is a current research challenge [8], [9].

In [10], the authors report that many image retrieval systems incorporating diversity have been applied to optimize the results and, consequently, improve the user satisfaction. These approaches generally rely on two phases: i) traditional retrieval of potentially highly relevant images; and ii) reranking these images to diversify. Moreover, most approaches that encompass diversity promotion apply clustering techniques in the second step.

Clustering is an unsupervised machine learning approach which aims to find natural groups of a set of samples, points, or objects. These groups may differ in terms of the form, size, and density. In general the optimal configuration for the set of points or objects is supposed to be compact and isolated. The clustering is used in order to provide a different view of the data, identifying degrees of similarity from the organization and summarization of these data [11].

These techniques are used to cluster the data into groups based on the similarity between them according to a predetermined criterion. Each of these groups may represent a query concept implying images with similar visual and/or textual features. The whole set of groups define a structure involving the distinct concepts related to the query. Hence, the diversity-oriented final ranking can be achieved, for instance, selecting a representative image from each conceptual group in a round-robin fashion. This selection aims at yielding a ranking semantically composed with images that convey coverage and novelty regarding the user-defined query and its underlying aspects.

There are many clustering algorithms available in the literature and, although there is no clear definition about the classification of these algorithms, they are generally divided into two categories [12]: partitioning methods and hierarchical methods. The partitioning methods are the most simple and are widely used in image retrieval. In turn, the hierarchical clustering methods are important when we want to organize the data within groups in different levels as a hierarchy. The hierarchical methods were broadly studied in this work due to their promising results in recent works [12], [13].

Despite the simplicity and promising results of the hierarchical clustering methods, some difficulties are encountered regarding the selection of the branches, that are the agglomeration points. This aspect is critical because a bad choice generates low quality clusters in terms of semantic and consistency. Results like this bring into focus current challenges and research directions, as the need to look for a tighter integration between the clustering algorithms and the application domain. In this sense, this integration can be achieved with a better understanding of the properties from the domain and the underlying structures of the applied algorithms [11].

In this paper, image descriptors and clustering algorithms are studied in order to reach a broader perspective of their individual and integrated effectiveness in the diversification process. First, we explore different filtering scenarios to analyze the behavior and effectiveness of the descriptors and clustering algorithms. Then, we apply a ranking criterion to compare the performances in both integrated and isolated views. The integrated view refers to the clustering results and the isolated view refers to the individual effectiveness of descriptors and algorithms. Finally, we discuss the implications of our work and the directions to tackle some challenges stated in the literature and also the ones highlighted by our experimental analysis.

II. Related Work

Image retrieval systems must find and retrieve in an effective manner a set of relevant images that provide, at the same time, some degree of diversity. In this section we discuss some works on image retrieval and visual diversification. We focus on works involving the application of clustering methods aiming to obtain good retrieval results considering the aspects of relevance and diversity.

In this context, the Retrieving Diverse Social Images [14], organized as part of MediaEval Benchmark [1], provided a database with images retrieved from Flickr [2], wherein each image is associated with a particular location or touristic attraction. The provided image database is divided into a development set (*devset*) and a test set (*testset*). Figure 1 presents images from the Arch of Triumph locality, exemplifying the images from the collection. Figure 1a shows possible query images for the locality. The challenge on this campaign is to refine the relevance-oriented results (1b) for each locality, selecting a subset that is, at the same time, relevant and diverse (1c).

The authors in [16] discuss the results achieved during the MediaEval 2013 campaign providing insights regarding the diversity task. Despite the advances reported by the works in the Retrieving Diverse Social Images task, there are much more room for improvement. The task comprises the relevance and diversity aspects, where the results should have representative relevant images with a maximum coverage and novelty of concepts regarding the query.

As a solution for the described task, the authors in [13] and [17] proposed a method that was divided into four steps: i) pre-filtering; ii) hierarchical clustering; iii) tree refining; and iv) result reranking. In the first step, the objective was to remove images considered as irrelevant [8]. To accomplish

[1] http://www.multimediaeval.org
[2] http://www.flickr.com

(a) Query images.

(b) Relevance-oriented result.

(c) Diversity-oriented result.

Figure 1: Example of search results illustrating relevance and diversity criteria. Source: [15].

such objective the authors used: face detection to remove images that contains people as the main subject[3]; geographic information to remove images that were shot far away from the queried location; user-credibility information[4] for removing images with a small number of views on Flickr; and an image focus measure to remove images out-of-focus or blurred, which were also considered as irrelevant for the task. In the following step, an image tree was constructed using the clustering algorithm, and then this tree was refined to improve the clusters final result, and, consequently, the diversification quality. For the construction of the hierarchy the Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) [18] algorithm was used. The BIRCH algorithm groups similar images in the same cluster or branch with the concept of clustering feature, which is a three dimensional vector that summarizes the cluster information. This vector is composed by the number of images in the cluster and by the linear and quadratic sums of the their visual features. The BIRCH clustering process is described in Algorithm 1.

The BIRCH algorithm starts examining all the images from the dataset, representing clusters with the clustering features, as stated in line 2. As indicated in line 3, the tree is built by these clustering feature vectors in an hierarchical structure using a centroid-based distance. When the tree is completely built, the second phase of the algorithm begins with a successive grouping of the vectors (line 5) using an agglomerative algorithm with, for instance, based on average inter-cluster distance. It is also at this point that the refinement is done, removing clusters that do not approach any other according to their distances. This whole process is oriented by two parameters, the branching factor that represents the number of entries per clustering feature, and the threshold that refers to the diameter that the leaf nodes must satisfy. Therefore, sparse clusters are removed and dense clusters are grouped into larger groups [12].

Algorithm 1: BIRCH

```
1 begin
2     Summarize clusters in clustering features;
3     Build the hierarchical tree based on clustering
        features;
4     repeat
5         Agglomerate clusters with selected clustering
            algorithm;
6         Remove clusters with low-density;
7     until Stopping criterion of algorithm to be attended;
8     return Refined clustering
9 end
```

With the final refined clustering, the diversified ranking was obtained by sorting the clusters in decreasing order based on the number of images, and selecting the most relevant image from each one. This selection occurs by choosing the first image with the highest score from each cluster. The next images to be chosen from the same clusters were selected based on the greater dissimilarity regarding the images previously selected. The effectiveness results demonstrated the consistency and stability of the proposed method. Nevertheless, their best results achieved approximately 0.45 of Cluster Recall [19], a metric that measures the percentage of clusters represented in the final ranking, which is still far from the optimum representativeness.

[3]This was an explicit constraint for the campaign.

[4]This data reflects the quality of tag-image content relationships.

Another method, proposed in [20], formalized the task as an approximate optimization problem based on [21]. The proposed methods used alternative formal definitions for relevance and diversity according to authors perspective from the task. For relevance, a supervised classification model was used to distinguish relevant images from irrelevant images, replacing the original relevance definition by the probabilistic model output. For diversity, the concept was redefined as the dissimilarity between the most similar image pair from a given set.

The state-of-the-art current results presented in [22] have achieved approximately 0.5 of Cluster Recall. This result reflects that this challenging context still needs to be further studied. In the context of the Retrieving Diverse Social Images task, the methods have been evaluated only based on the final diverse ranking, while the quality of the groups that were taken as input for that final ranking was not actually assessed.

In our work, we highlight the fact that this grouping effectiveness is actually an important step for diversification since the reranking step directly relies on it. Moreover, for the construction of the final ranking, multiple alternative methods may be applied, e.g., in regard to intra and inter-cluster sorting and representative image selection. At the same time, considering currently applied evaluation measures, such as Precision and Cluster Recall, in regard to a given partition of the ground-truth, no matter what image is taken as representative, the final numerical effectiveness do not change. Hence, as a natural assumption, the closer are the clusters constructed to the real partitions, the more beneficial it is for the reranking step taken afterwards. Therefore, instead of evaluating diversity as a property of the final ranking, we propose taking a step backward and analyzing the actual effectiveness of algorithms and features to properly map the input data to conceptual groups.

III. Proposed Analysis

In this paper we intend to understand the real capability of features and algorithms to reproduce what is performed by a human being, organizing images into conceptual groups. Hence, we propose to evaluate several descriptors and algorithms regarding their clustering effectiveness in comparison to the partitions generated by real users.

Initially, we considered three different scenarios regarding relevance-based filtering of the results in order to perform the clustering effectiveness evaluation. Hence, results were analyzed in a real scenario wherein no filtering is used; in an ideal scenario with noisy (non-relevant) information filtered out; and in a guided scenario wherein, besides noise filtering, the target number of groups for each test query is also known a priori. Therefore, for the different scenarios, we exploit different dataset configurations: a dataset version containing relevant and non-relevant images in the real scenario, and another version composed only by relevant images in both ideal and guided scenarios.

In this first analysis, we intend to construct a broad understanding of important aspects regarding clustering results for diversity: the effect of noise in the dataset, the impact of the number of clusters, and the overall clustering quality. For the first aspect, we intend to understand how the non-relevant images are tackled in the clustering process for diversity, and to what extent they impact the retrieval results. With respect to the number of clusters, the literature claims that this choice is a hard task and particular to the data [11]. Moreover, the clustering associates similar data regarding a concept, which is itself a subjective and user-related aspect. Therefore, analyzing the performance of the clustering algorithms when the number of concepts are known a priori is an interesting and important study for assessing whether it is a limiting factor or not. Finally, the proper clustering quality assessment in regard to the representativeness of the multiple possible concepts is quite important since the accurate detection of such concepts could provide extrinsic improvements for the diversification objective.

From the clustering performance in different scenarios, we conducted a deeper analysis to assess the individual performances of descriptors and algorithms. For this analysis, we propose three ranking criteria related to descriptors and algorithms effectiveness: the matching measures, which quantifies how close are the clusters in comparison to the partitions generated by human beings; the relative ranking positions, which establish the average ranking position of descriptors and algorithms in terms of the matching measures; and the impact of filtering on their clustering effectiveness. In this second step, we aimed at highlighting features and algorithms promising to the application domain, as well as a tighter integration for the diversity objective.

IV. Experimental Setup

In this section we describe the settings for the experimental effectiveness evaluation. These aspects refer to the database, visual properties, text similarity measures, clustering validation measures.

A. Dataset

The descriptors and algorithms were evaluated using the image collection from the Retrieving Diverse Social Images Task [14] from MediaEval 2015. We use the devset to perform the experiments. The *devset* includes 153 localities wherein each locality there are roughly 300 images obtained from Flickr. Moreover, each locality also contains additional information, such as visual descriptors, textual metadata, and relevance and diversity ground-truth.

The ground-truth is an important aspect in this work since it was used to conduct the clustering quality evaluation. As described in [14], this ground-truth was generated from real specialist user annotations. For relevance, the users were required to analyze each image for annotation as relevant or non-relevant. In this process, it was allowed to use additional resources, as the Internet, to help them in the decision-making. For diversity, the users were required to group the relevant images with similar visual concepts in clusters and label the clusters with appropriated names. For the final groups definition, all the process of ground-truth generation was done with voting and revision procedures. For relevance assessment

of the *devset*, 11 human assessors were involved. In turn, for diversity ground-truth generation, the devset was labeled by 3 individuals, which annotated different parts of the data by visually grouping the images.

B. Visual Properties and Textual Similarity

We selected 38 features for analysis, including visual descriptors and text measures. Besides the descriptors provided by the Retrieving Diverse Social Images task, we used all descriptors reported in [23].

1) Visual Descriptors: Although visual descriptors can extract multiple information, they can be roughly divided according to the main visual aspect represented. Therefore, considering it, we can group the evaluated descriptors into the following categories:

- Color: ACC [24], BIC [24], [25], CM, CM3x3, CN, CN3x3, CLD, CSD, JCH, LUM, OPHIST, SCD, SCH;
- Texture: CEDD, EHD, FCTH, Gabor, GIST [26], GLRLM, GLRLM3x3, JCD, LAS [24], LBP, LBP3x3, PHOG, and Tamura;
- Structure: CNN_AD, CNN_GEN, DSM, HOG, HSM, HSWSA, WSA.

2) Text Similarity Measures: The text measures were provided by the task were the TF, DF, and TF-IDF. The TF is the term frequency that corresponds to the number of occurrences in the text field, the DF is the document frequency representing the number of entities with the term, and the TF-IDF is the ratio between TF and DF [14]. Moreover, the following textual descriptors were extracted: Cosine [27], BM25 [27], Dice [28], and Jaccard [28].

C. Clustering Algorithms

For the proposed experiments, we evaluated six clustering algorithms: one partitional and five hierarchical. We mainly focus on the hierarchical algorithms taking into account an intrinsic relation with diversity in terms of concepts hierarchies and for the promising results that have been reported on different domains [12]. As partitional algorithms have been extensively applied on recent works, we intend to evaluate the potential of alternative hierarchical methods. Moreover, to the best of our knowledge, no previous work have comprehensively and comparatively evaluated the effectiveness of partitional and hierarchical methods for the diversity promotion task.

The hierarchical algorithms evaluated are: single-link, complete-link, average-link [29], BIRCH [18][5], and Chameleon [30]. The three first are classic algorithms commonly used to create groups with hierarchies. As an important parameter, most of them take the number of clusters as a stopping criterion. In our experiments, based on previous empirical results, it was set to 40 clusters. Finally, the last two are promising hierarchical algorithms [12]. They do not need the number of clusters as parameter. Notably, to the best of our knowledge, the Chameleon algorithm was not previously used for diversity promotion purposes.

Considering our proposal on evaluating the Chameleon algorithm for diversity promotion, we present here some details regarding its procedure. Chameleon clustering process is composed by two phases, which are presented in Algorithm 2. In the first phase, a graph of the collection is built with the k-nearest neighbours criterion (line 2). This graph is partitioned for the generaton of subclusters using the METIS' multilevel *k*-way partitioning [30] (line 3). In the second phase, agglomerations are performed based on the relative interconnectivity and relative closeness of the subclusters (line 5). These criteria are responsible for the dynamic modeling methodology of the algorithm. In [30], the authors claim the possibility of obtaining high quality clusters regardless of their forms.

Algorithm 2: Chameleon Clustering Algorithm

```
1 begin
2     Build graph with the k-nearest neighbours;
3     Partition the graph in subclusters;
4     repeat
5         Find the the most interconnected and close
            clusters;
6         Agglomerate them;
7     until Maximize interconnectivity and closeness
        threshold;
8     return Clusters
9 end
```

Finally, the partition-based algorithm evaluated was the well-known k-Medoids [12], which was taken as baseline for its successful application in previous works [23].

D. Clustering Assessment

Given the context of clustering as a diversification approach, one important aspect to be evaluated is whether the clustering structures discovered are satisfactory. To assess the clustering results, we may apply several evaluation measures to analyze the quality of the clusters generated. To this assessment, we use the term cluster in regard to the groups generated by clustering algorithm and partitions in regard to groups manually generated by human beings. We use extrinsic matching based measures considering that the ground-truth is known. The matching character means that the measure quantify the extent to which a cluster contains images from a certain partition of the ground-truth.

As described in Section IV-A, the ground-truth provided was generated by real specialist users. In particular, we compute the following measures: purity, maximum matching and F-measure [31]. These measures convey values between 0 and 1, wherein 0 indicates no similarity between clusters and partitions, and 1 indicates an exact matching.

The purity measure (Equation 1), quantifies to what degree a cluster c_i has objects from a ground-truth partition t_j.

$$purity_i = \frac{1}{n_i} \max_{j=1}^{k} \{\Omega_{ij}\} \quad (1)$$

[5]For runtime purposes and using pre-computed similarity measures, we have adapted the BIRCH algorithm on the phase where centroids are calculated, replacing it by a medoids definition process.

where n is the number of objects from the clustering and Ω_{ij} is the number of common objects between a cluster and a partition. For each cluster, the purity measure looks for the partition to which the cluster has the greatest number of common objects. Thus, the measure quantifies how pure a cluster is based on the ratio between the intersection of objects with that partition and the total number of objects from the particular cluster.

To assess the purity for the entire clustering one just need to perform the sum of individual purities, as shown in Equation 2.

$$purity = \frac{1}{n}\sum_{i=1}^{r}\max_{j=1}^{k}\{\Omega_{ij}\} \tag{2}$$

By definition, the purity measure allows more than one group to have the greatest intersection with the same partition. Alternatively, the maximum matching measure deals with this aspect while aims to maximize the sum of common objects between clusters and partitions. However, only one cluster can match with a given partition. The maximum matching measure has a graph-based definition, wherein the clusters and partitions are taken as vertices and the links, representing the existence of objects in common, are the edges. This measure is formulated as shown in Equation 3.

$$match = \arg\max_{M}\{\frac{w(M)}{n}\} \tag{3}$$

where M is the subset of edges that do not have common vertices and $w(M)$ is the sum of weights of the edges from M.

As an illustration, Figure 2 presents a clustering configuration for the K-Means algorithm where the partitions from the ground-truth are known. Each cluster is represented by different symbols (squares, triangles, and circles). The centroids are represented by the black objects, the objects with gray color are in the correct partition and the white ones were clustered in a wrong place.

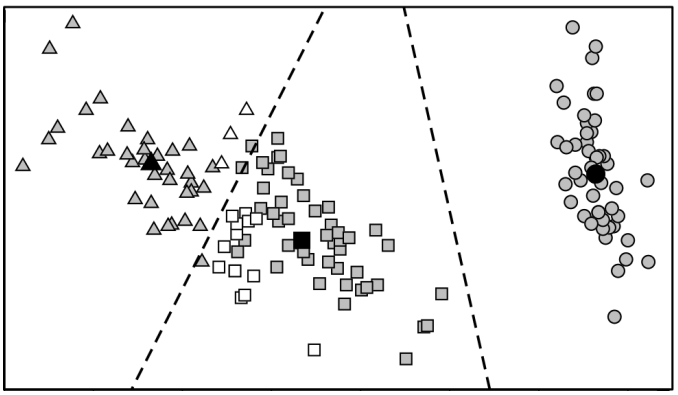

Figure 2: Clustering example with available ground-truth presenting a high matching degree. Source: [31].

Observing the clustering from Figure 2, we can notice that it presents a high purity degree taking into account the number of objects in common with the respective partition, and the low number of objects erroneously grouped. For this example, the maximum matching measure provides the same result due to the clusters matching are already on the configuration that maximize the weighted sum.

From Figure 3, we can notice a contrast regarding the result on Figure 2. Besides the number of objects grouped in a wrong manner being relatively high as we can see looking at the white triangles, the matching of each cluster with a unique partition is not attended for the square and circle clusters considering the gray color meaning. These clusters share the same partition regarding the maximum number of objects, thus, the maximum matching value reflect a penalization in this configuration, indicating a low quality of the clustering.

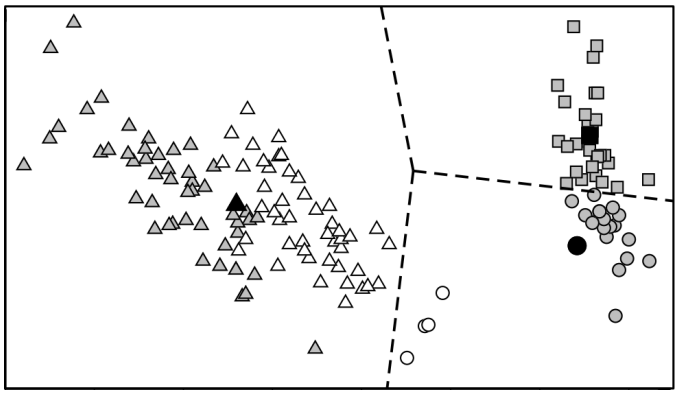

Figure 3: Clustering example with available ground-truth presenting a low matching degree. Source: [31].

Finally, we also considered the F-measure, which is defined as the harmonic mean between Precision (Equation 4) and Recall (Equation 5). The precision (P) of a cluster represents the ratio between the cluster objects in common with the partition with the maximum overlap and the total number of objects from the cluster, similar to purity. The recall (R) of a cluster represents the common objects of a partition regarding the corresponding cluster with the maximum overlap.

$$P_i = \frac{1}{n_i}\max_{j=1}^{k}\{\Omega_{ij}\} \tag{4}$$

$$R_i = \frac{1}{m_j}\max_{j=1}^{k}\{\Omega_{ij}\} \tag{5}$$

where n_i is the number of objects from the cluster i, Ω_{ij} is the common objects between a cluster i and a partition j, and m_j is the number of objects from the partition j. The F-measure for a cluster is given by Equation 6, which may be averaged for all clusters.

$$F_i = \frac{2 \times P_i \times R_i}{P_i + R_i} \tag{6}$$

The F-measure for the entire clustering is simply the average of individual F-measures from each cluster. The maximum value that can be obtained, the value 1, indicates an ideal clustering.

V. Results and Discussion

We have conducted the following experimental evaluations: Section V-A presents the general clustering evaluation considering the three filtering scenarios and Sections V-B and V-C discuss the clustering effectiveness considering the comparative of descriptors and algorithms.

A. Clustering Performance Analyses

The different filtering scenarios studied intended to allow an evaluation of the algorithms and descriptors considering the measures of purity, maximum matching and F-measure. The clustering algorithms were ran in combination to each of the 38 descriptors for all 153 localities. Each combination was evaluated with the matching measures for each locality and for the different filtering scenarios. In total, 228 combinations of algorithms and descriptors were executed for each scenario, except for the guided scenario for which we executed 152 combinations due to BIRCH and Chameleon not taking the number of clusters as input. For each combination, all individual matching measures were computed. In this section, for analysis purposes, we report and discuss only purity results for the complete-link algorithm due to the correlation and similar behavior observed in the results for all algorithms and measures. The experimental results are discussed in the following subsections.

1) Real Scenario: The real scenario represents a dataset containing relevant image as well as non-relevant images. When we considered the use of no relevance-based filtering, we observed that the best results achieved a maximum value around 0.3 for the purity measure. For instance, Figure 4 presents the purity-based evaluation for the complete-link algorithm.

As we can see, the best feature achieved only 0.31 and, for the same algorithm and metric, some features fall close to 0.1. These results reflect the difference between the generated clusters and the real partitions in the dataset and indicates an noticeable limitation of the algorithms to deal with non-relevant or noisy images. For the other metrics, the values indicated even smaller correspondence to the human-generated partitions.

2) Ideal Scenario: With the objective o assessing the isolated performance of the clustering, we applied a perfect filtering approach to adjust the dataset to contain only relevant images. From this scenario, an improvement in the results was expected. The experiments supported this as it is shown in Figure 5 for the same algorithm discussed in the previous section.

The complete-link algorithm achieved a result of approximately 0.53 of purity and, for all combinations, there were significant gains in comparison to the real scenario. However, even in an ideal dataset, the best paring of clustering algorithms and features only represents clusters in roughly half of the ideal purity degree of similarity considering the concepts separation from real users. Moreover, at this point, we may observe a trend related to the existence of a common subset of descriptors with the best values independent of metric, algorithm, and even the scenario. We may highlight the promising clustering performance of BIC, BM25, DICE, DSM, HSM, and Jaccard.

3) Guided Scenario: According to [11], automatically determining the number of clusters has been one of the most difficult problems in the clustering task. Considering this, we assume a different scenario beyond only taking relevant images. In this scenario, the number of clusters is also assumed as known a priori. We intend to understand how this aspect could impact the clustering performance. Following the previous results for complete-link purity, Figure 6 shows the results for this guided scenario.

Despite the exact tuning on the number of clusters, the complete-link algorithm did not present significant improvements. Similarly, the other algorithms and metrics presented the same behavior and no significant gains. Therefore, we believe the semantic gap problem is a harder limitation beyond the number of clusters. These results highlight the difficulty of the low-level features on providing the cluster algorithms with proper similarity evidences for the unsupervised feature space partitioning, which was not positively affected even when the number of partitions was known in advance.

B. Descriptors Performances Analyses

In this section, we discuss the comparative analysis conducted between descriptors. We intend to understand their effectiveness for image clustering in the target domain. The performance analyses for the descriptors are organized and discussed considering the following aspects: matching-based effectiveness, relative ranking positions, and filtering impact.

1) Matching-based Effectiveness Comparison: For this analysis, the 38 descriptors were compared using the three measures. In particular, the scope selected was the ideal scenario. We evaluated all the combinations of algorithms (6) and localities (153). Therefore, for each descriptor, we used 918 effectiveness measurements and we report average values and confidence intervals (95% confidence). Figure 7 presents the results sorted by average purity. The top performing descriptors were the HSM and DSM, which are Bag-of-Visual-Words-based descriptors. These descriptors are based on sparse (Harris-Laplace detector) SIFT, with 512 visual words (randomly selected), soft assignment ($\sigma = 150$), and max pooling. These descriptors also achieved the best result for maximum matching and had a high F-measure performance when compared to the other algorithms. Notably, the small confidence intervals shows stable results regarding the effectiveness of the descriptors. Besides that, these results demonstrate a correlation in terms of the highlighted descriptors from the analysis in the previous scenario. Moreover, in Figure 7 it is also possible to observer descriptors with low performances for the domain.

2) Relative Ranking Positions: Considering the 3 scenarios, for each of the 6 clustering algorithms, a ranking of descriptors was constructed based on their matching-based effectiveness results. We used 48 rankings[6] from the 3 matching measures to obtain the performances by average positions. Figure 8

[6]The number of rankings is smaller since BIRCH and Chameleon do not allow the guided scenario for not taking the number of clusters as input.

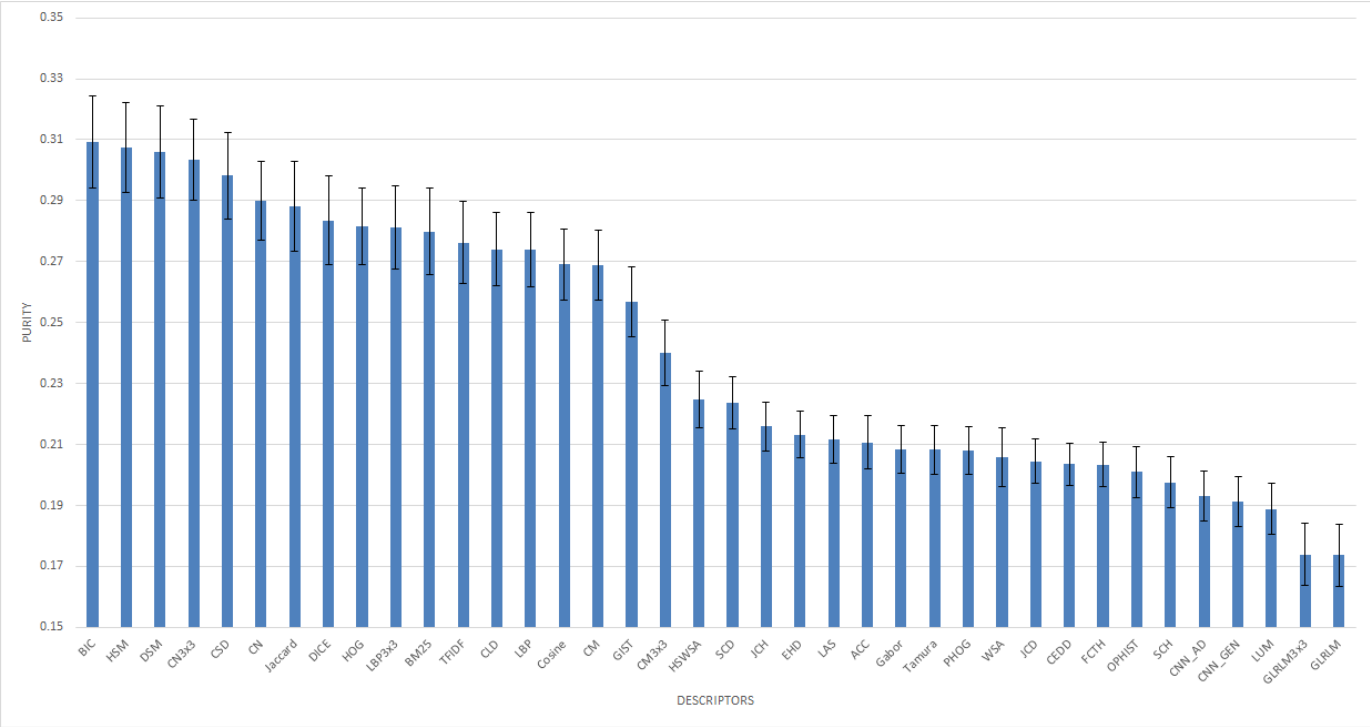

Figure 4: Experimental results for complete-link purity in the real scenario.

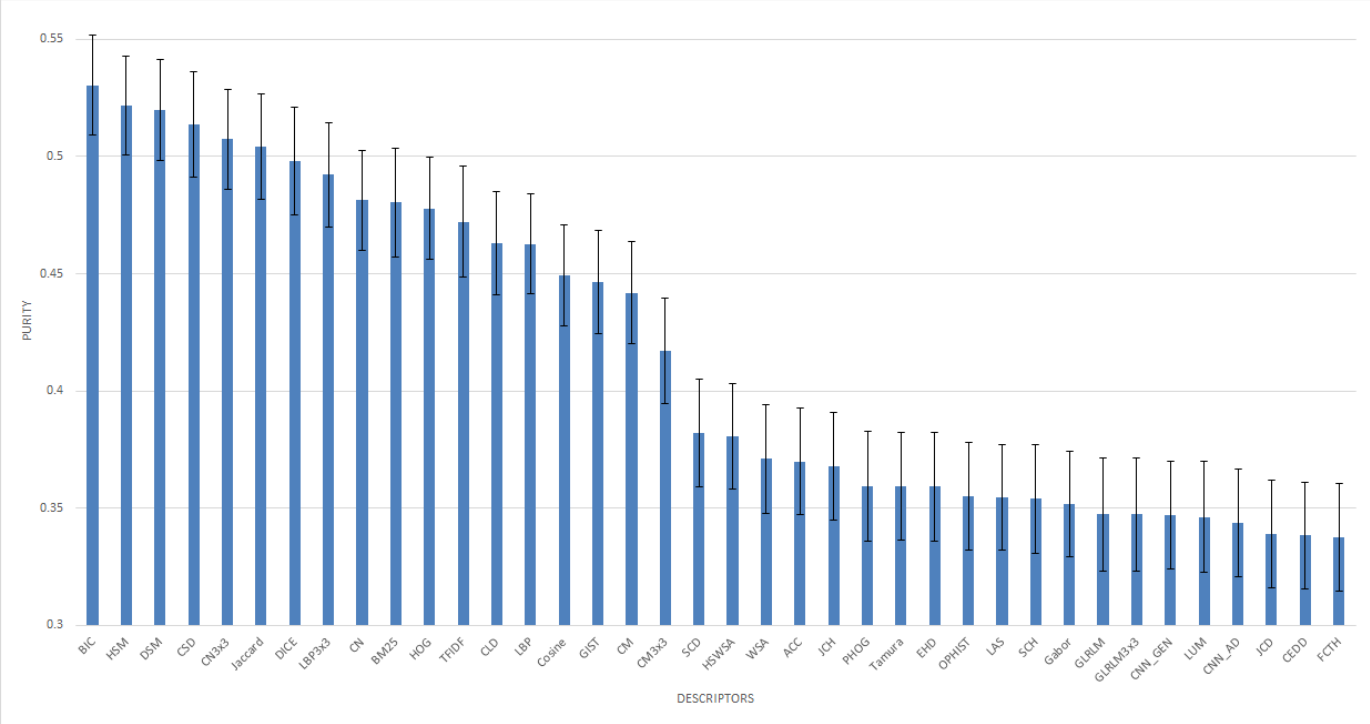

Figure 5: Experimental results for the complete-link purity in ideal scenario.

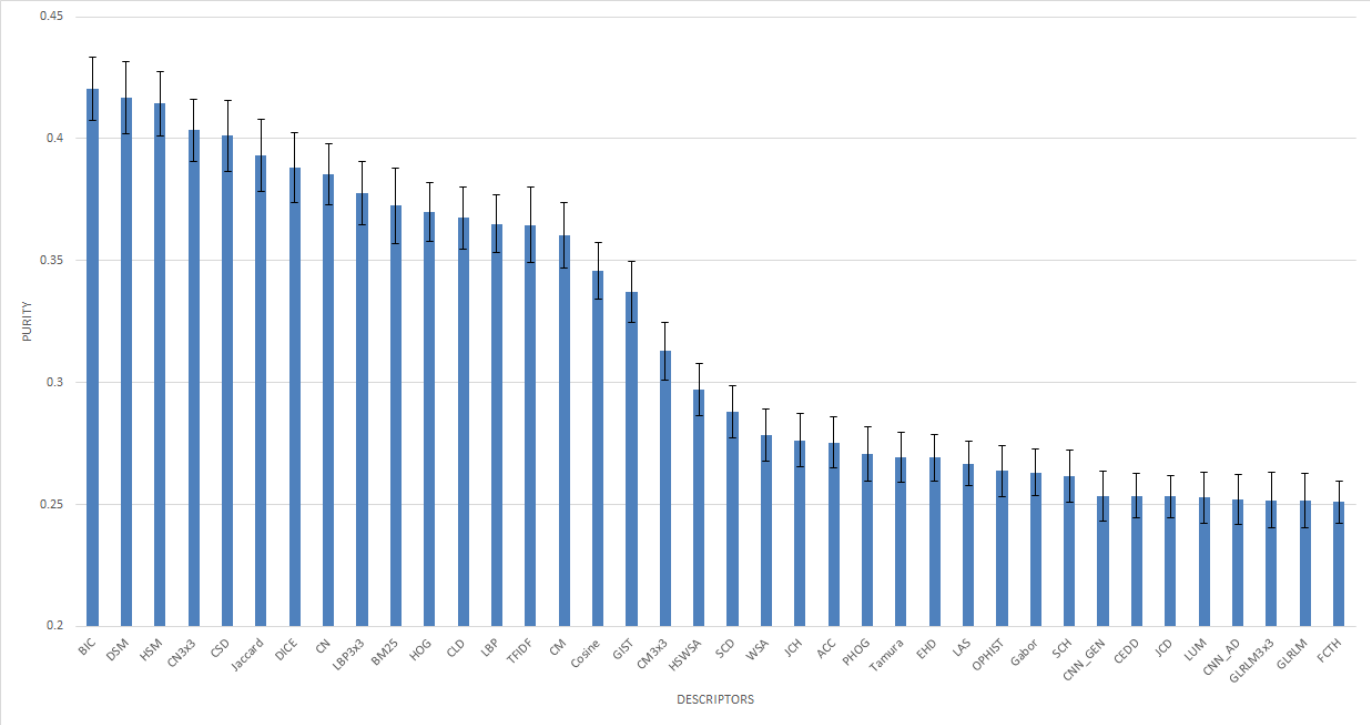

Figure 6: Experimental results for the complete-link purity in the guided scenario.

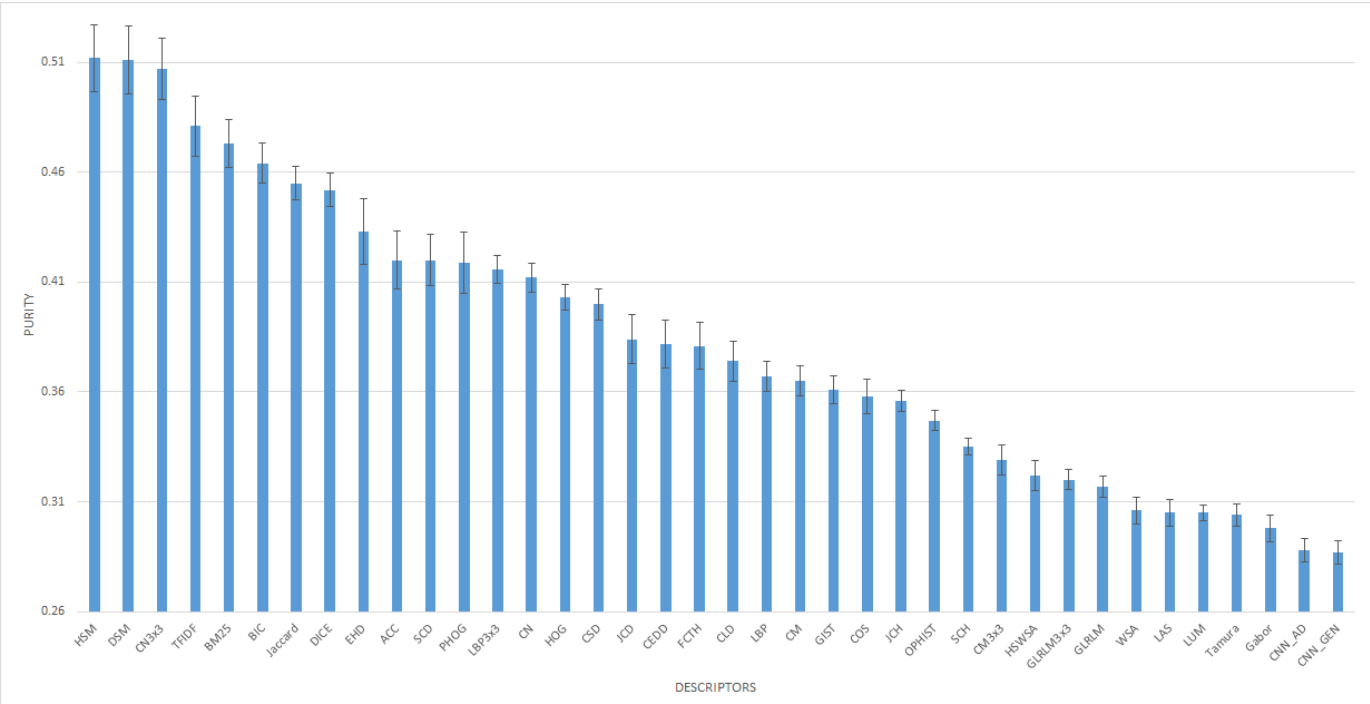

Figure 7: Descriptors sorted by average purity values.

presents the final average ranking results and confidence intervals (95% confidence) for all rankings.

From the relative ranking analysis, it was possible to notice common best descriptors for the different measures, as also stated in the previous discussion. Besides that, the relative ranking provides a broad perspective of the descriptors performances. For instance, the Jaccard descriptor did not have the best purity values considering its results for all scenarios, but it demonstrates stability and relative high performance with the best overall relative ranking. Besides, we have also noticed a similar behavior when separately taking the different filtering scenarios.

3) Filtering Impact: Finally, the descriptors were analyzed in terms of the filtering impact for each measure. We compared the measures for all descriptors regarding the relative gains from the real scenario to the ideal scenario. For instance, Figure 9 shows the relative gains for purity.

With theses results we can see the benefits of filtering and observe that the descriptors with lower values in terms of measures (Figure 7) and rankings (Figure 8) tend to have important relative gains. Therefore, these results suggest that the descriptors with the best average measures are more stable with regard to noise (non-relevant) items. Furthermore, an interesting fact is that the filtering step demonstrated to be a really important aspect since all descriptors achieved absolute gains.

C. Algorithms Performances Analyses

Similar to the evaluation conducted for the descriptors, we have also investigated the general effectiveness of the clustering methods studied. In the following, we present the results and discussions of the matching-based and relative ranking effectivenesses of the clustering algorithms.

1) Matching-based Effectiveness Comparison: For each algorithm, we obtained 5.814 clustering configurations for each measure considering the 153 localities and the 38 descriptors. The six clustering algorithms were evaluated by their average purity in the ideal scenario. Table I shows the average purities for the algorithms. In general, the BIRCH achieved the best results, confirming its promising effectiveness. Even so, the effectiveness of all algorithms are close for all metrics and the confidence intervals were considered too large, indicating a high variation in the results for all locations/scenarios and a high influence of the features encoded by the descriptors on the effectiveness.

2) Relative Ranking Positions: This ranking analysis was done by computing the total frequency of the best performing algorithm for all the descriptors. For example, considering the BIC descriptor, we selected the best instance taking into account its usage with the six algorithms. The ideal scenario was considered and this analysis was done for each measure. Considering the ranking position analysis, the algorithms presented a behavior similar to the previous analysis, where BIRCH had superior effectiveness as the best performing algorithm for most of the evaluated features (21) as presented in Table I.

Table I: Algorithms Purity, Number of Features They Best Perform and Relative Gain.

Algorithm	Average Purity	Best for n descriptors	Relative gain (%)
K-medoids	0.42 ± 0.02	$n = 11$	70.8
Single-link	0.37 ± 0.01	$n = 2$	95.0
Complete-link	0.42 ± 0.02	$n = 3$	72.3
Average-link	0.41 ± 0.02	$n = 2$	81.9
BIRCH	0.49 ± 0.09	$n = 21$	45.6
Chameleon	0.21 ± 0.00	$n = 0$	49.2

3) Filtering Impact: Finally, similar to the descriptors, regarding the filtering impact, the algorithms had relative gains on purity where, in general, the impact was higher the lower was the algorithm performance (Table I). An exception was the Chameleon algorithm, which had a lower filtering impact despite its matching-based effectiveness results. We believe that, in general, its effectiveness was reduced due to the characteristics of the relative thresholds used to agglomerate the objects and generate the clusters. This is an aspect also claimed by its authors in [30], where they state the possible necessity to adjust the relative concepts of interconnectivity and closeness according to the application domain. Despite these purity results, in general, the algorithms tend to present a uniform filtering impact. Overall, these results suggest that the filtering is decisive to the clustering algorithms effectiveness and highlights the difficulty that noisy items bring to the grouping process.

VI. Conclusions

In this study, we performed an experimental analyses targeting a broad perspective and understanding of the image clustering task for visual diversification. The experiments demonstrated a low similarity between automatically generated clusters and human-generated partitions regardless the filtering scenario applied. In this context, the descriptors effectiveness analyses demonstrated the importance of applying the proper features in regard to the data clustering domain. The experiments also reflects the fact that there is no better clustering algorithm, demonstrating the need to achieve a tighter integration between the algorithms and the application needs or even in a more profound way to specific queries or contexts.

Additionally, we highlighted the best descriptors for the domain based on the matching-based effectiveness comparison and relative ranking position analyses. In the first analysis, we have shown that the descriptors that best perform were roughly the same for all algorithms. In the second analysis, the previous results were corroborated since the descriptors with best average ranking positions were also roughly the same best performing descriptors from the previous analysis.

Furthermore, considering the filtering impact, the relative gain results show the importance of the noise removal step as all descriptors achieved superior performance when in the ideal scenario. Finally, we conclude that even the best performing descriptors for each clustering algorithm in an ideal filtering scenario are still distant from the human being conceptual understanding. These findings ratifies the semantic gap problem as the main limiting aspect and highlights the need for further development.

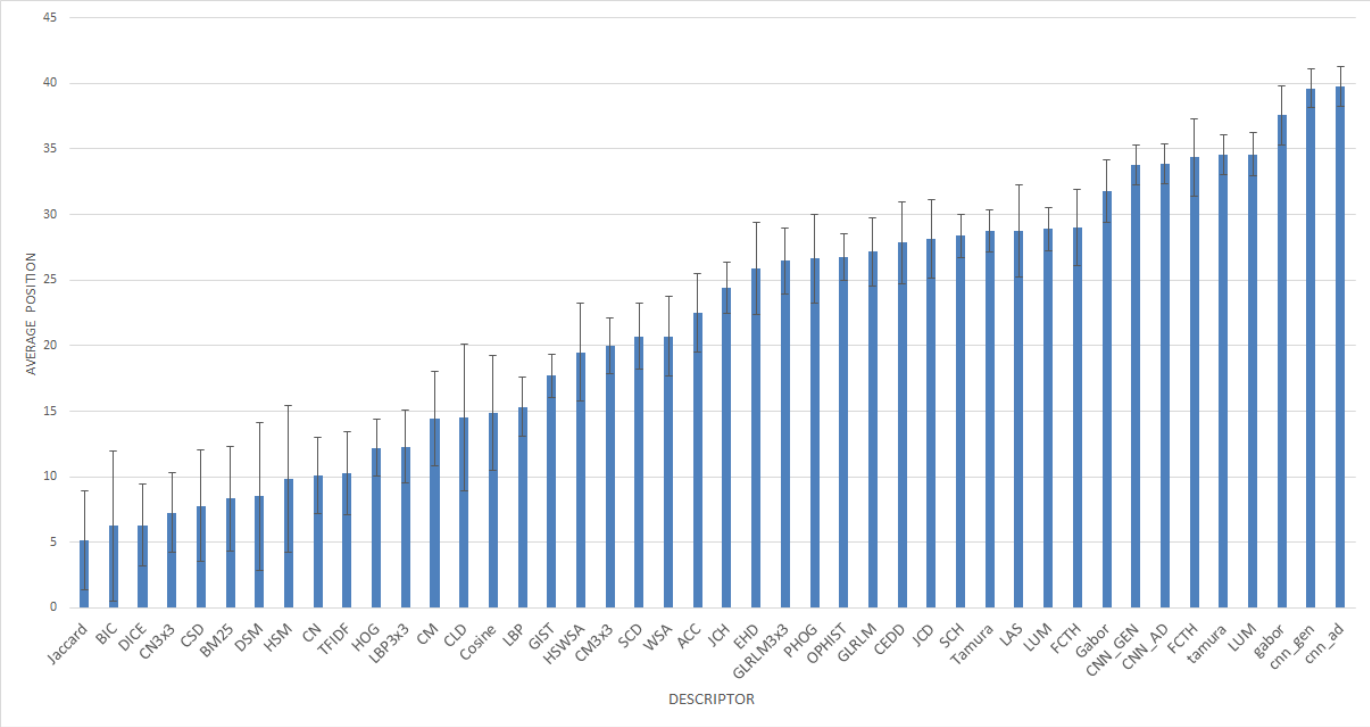

Figure 8: Descriptors sorted by average positions.

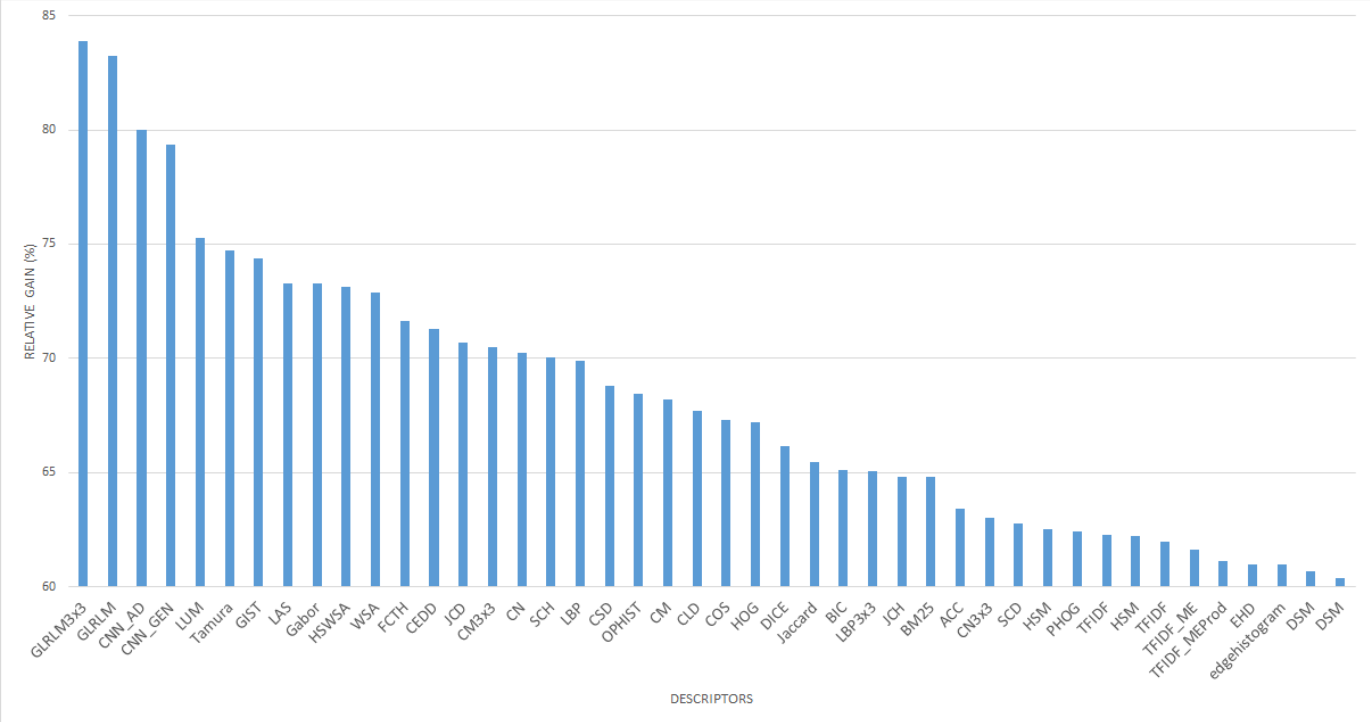

Figure 9: Descriptors relative gains for purity between the real and ideal scenarios.

VII. Future Work

As future work, as an attempt to attenuate the semantic gap, we will evaluate a multimodal feature approach to improve the low-level features representation of images. Moreover, regarding the performance of the algorithms, we may apply ensemble clustering techniques to obtain better decisions and, consequently, better clusters. Thus, we intend to use a co-operative approach to build a more mature instance. With this, we hope to overcome the semantic gap in some extent considering subjective and objective aspects related to the task. Finally, important effectiveness advances may be achieved by introducing dynamic processing in terms of features and algorithms for different retrieval scenarios and queries.

References

[1] R. da Silva Torres and A. X. Falcão, "Content-based image retrieval: Theory and applications." *RITA*, vol. 13, no. 2, pp. 161–185, 2006.

[2] Y. Rui, T. S. Huang, and S.-F. Chang, "Image retrieval: Current techniques, promising directions, and open issues," *Journal of Visual Communication and Image Representation*, vol. 10, no. 1, pp. 39–62, 1999.

[3] R. C. Veltkamp and M. Tanase, *Content-Based Image and Video Retrieval*, 2002, ch. A Survey of Content-Based Image Retrieval Systems, pp. 47–101.

[4] I. Ounis, C. Macdonald, and R. L. Santos, "Search result diversification," *Foundations and Trends in Information Retrieval*, vol. 9, no. 1, pp. 1–90, 2015.

[5] R. L. T. Santos, C. Macdonald, and I. Ounis, "Search result diversification," *Foundations and Trends in Information Retrieval*, vol. 9, no. 1, pp. 1–90, Mar. 2015.

[6] R. T. Calumby, R. da Silva Torres, and M. A. Goncalves, "Diversity-driven learning for multimodal image retrieval with relevance feedback," in *IEEE International Conference on Image Processing*, 2014, pp. 2197–2201.

[7] R. L. Santos, C. Macdonald, and I. Ounis, "Intent-aware search result diversification," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011, pp. 595–604.

[8] B. Ionescu, A. Popescu, M. Lupu, A. L. Gınsca, and H. Müller, "Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation," in *MediaEval 2014 Workshop, Barcelona, Spain*, 2014.

[9] M. Drosou and E. Pitoura, "Multiple radii disc diversity: Result diversification based on dissimilarity and coverage," *ACM Transactions on Database Systems*, vol. 40, no. 1, p. 4, 2015.

[10] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol, "Visual diversification of image search results," in *Proceedings of the 18th ACM International Conference on World Wide Web*, 2009, pp. 341–350.

[11] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[12] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and techniques*. Morgan Kaufmann, 2012.

[13] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F. De Natale, "Retrieval of diverse images by pre-filtering and hierarchical clustering," *Working Notes of MediaEval*, 2014.

[14] B. Ionescu, A. Popescu, M. Lupu, A.-L. Gînsca, and B.-B. Müller, Henning, "Retrieving diverse social images at mediaeval 2015: Challenge, dataset and evaluation." in *Working Notes of MediaEval*, 2015.

[15] F. S. Cordeiro, V. P. Santana, and R. T. Calumby, "Diversify - um micro arcabouço para avaliação de métodos de sumarização visual," *XVIII Seminário de Iniciação Científica da UEFS*, 2014.

[16] B. Ionescu, A. Popescu, A.-L. Radu, and H. Müller, "Result diversification in social image retrieval: a benchmarking framework," *Foundations and Trends in Information Retrieval*, vol. 9, no. 1, pp. 1–90, 2015.

[17] D.-T. Dang-Nguyen, G. Boato, F. G. Natale, L. Piras, G. Giacinco, F. Tuveri, and M. Angioni, "Multimodal-based diversified summarization in social image retrieval," *Working Notes of MediaEval*, 2015.

[18] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: an efficient data clustering method for very large databases," in *ACM SIGMOD Record*, vol. 25, no. 2, 1996, pp. 103–114.

[19] C. X. Zhai, W. W. Cohen, and J. Lafferty, "Beyond independent relevance: methods and evaluation metrics for subtopic retrieval," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 10–17.

[20] E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, and I. Vlahavas, "Socialsensor: Finding diverse images at mediaeval 2014," *Working Notes of MediaEval*, 2014.

[21] T. Deselaers, T. Gass, P. Dreuw, and H. Ney, "Jointly optimising relevance and diversity in image retrieval," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009, p. 39.

[22] S. Sabetghadam, J. Palotti, N. Rekabsaz, M. Lupu, and A. Hanburry, "Tuw @ mediaeval 2015 retrieving diverse social images," *Working Notes of MediaEval*, 2015.

[23] R. T. Calumby, I. B. A. d. C. Araujo, V. P. Santana, J. A. Munoz, O. A. Penatti, L. T. Li, J. Almeida, G. Chiachia, M. A. Gonçalves, and R. d. S. Torres, "Recod @ mediaeval 2015: Diverse social images retrieval," *Working Notes of MediaEval*, 2015.

[24] O. A. Penatti, E. Valle, and R. d. S. Torres, "Comparative study of global color and texture descriptors for web image retrieval," *Journal of Visual Communication and Image Representation*, vol. 23, no. 2, pp. 359–380, 2012.

[25] R. O. Stehling, M. A. Nascimento, and A. X. Falcão, "A compact and efficient image retrieval approach based on border/interior pixel classification," in *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2002, pp. 102–109.

[26] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 300–312, 2007.

[27] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England, 2011.

[28] J. Lewis, S. Ossowski, J. Hicks, M. Errami, and H. R. Garner, "Text similarity: an alternative way to search medline," *Bioinformatics*, vol. 22, no. 18, pp. 2298–2304, 2006.

[29] G. Gan, C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*. Siam, 2007, vol. 20.

[30] G. Karypis, E.-H. Han, and V. Kumar, "Chameleon: Hierarchical clustering using dynamic modeling," *Computer*, vol. 32, no. 8, pp. 68–75, 1999.

[31] M. J. Zaki and W. Meira Jr, *Data Mining and Analysis: Fundamental Concepts and Algorithms*. New York, NY: Cambridge University Press, 2014.