Features Fusion for Diversity Gap Reduction

Iago Breno Alves do Carmo Araujo[1], **Rodrigo Tripodi Calumby**[1]

[1]University of Feira de Santana – Feira de Santana – BA – Brazil

{ibacaraujo, rtcalumby}@ecomp.uefs.br

Abstract. *Diversity has been promoted in image retrieval results using clustering algorithms to tackle queries, which refer to multiple information needs, e.g., due to ambiguity. Despite the effective results of diversity-aware methods, the image wealth of large collections and the subjectivity of human perception bring the semantic gap problem. This paper presents multimodal fusion approaches aimed at reducing the diversity gap with ensemble clustering and dimensionality reduction. The applied methods were evaluated by quantifying the clustering effectiveness in comparison to human decisions. The experimental results demonstrate the potential of these approaches to boost diversity-oriented engines and that they could improve state-of-the-art systems.*

1. Introduction

With the advances in data storage and image acquisition, the design of image search engines became a challenging task. Multiple approaches were developed for tackling user needs by providing effectiveness on the management of large multimedia datasets. For instance, content-based image retrieval (CBIR) systems are broadly used for similarity-based search based on visual features. However, similarity-based results have been shown as insufficient to satisfy the actual information need [Ounis et al. 2015]. The origin usually lies on the multiplicity of query intents. Quite frequently, a textual query can express multiple information needs. This ambiguity may lead a system to wrong results. Alternatively, providing a query with an example image may help attenuating such problem. Although it is effective for generating non-ambiguous results, the system could retrieve several near-duplicate images. Therefore, the results would be deteriorated with redundancy. The ambiguity and redundancy issues motivated the development of diversity-oriented retrieval approaches [Ounis et al. 2015]. In CBIR, a diverse image search result comprises a set of images relevant to the query but with complementary visual semantics [Ionescu et al. 2015]. Therefore, many diversification approaches rely on clustering techniques to achieve diversity [Boteanu et al. 2015] by selecting representative images from the discovered clusters.

Despite the result improvements and information gain achieved, the CBIR methods require the extraction of low-level descriptions, quite frequently not capturing the high-level semantics of the images. The extracted features represent visual characteristics, such as color and texture, which inherently carry a limitation due to the image visual wealth and the rich semantics of human perceptions, which defines the semantic gap problem [Veltkamp and Tanase 2002]. Moreover, CBIR systems propagate the semantic gap throughout the retrieval process and the similarity-based diversification inherits such problem. In the clustering task, human beings are able to visually group the images based on a rich understanding, while clustering algorithms using low-level features have limited

perception of such aspects and frequently fail the task. Hence, the distance between the automatic clustering and the manual grouping from humans is what we define here as the *diversity gap*, given it has a direct impact on the final diverse ranking.

Data fusion has been shown as effective for attenuating the semantic gap and has also been applied to diversification [Liang et al. 2014]. This paper focus on the intrinsic relationship between the clustering quality and the diversity of the ranking. Therefore, we apply two unsupervised fusion approaches and evaluate their potential for diversity enhancement, which consider important aspects, such as the input data and the objective function. The first approach relies on the fusion of multiple clusterings for combining the partitioning discovered with different features. Hence, it allows combining multiple views of the data into a unique structuring. The second approach aims at creating enhanced low-level representations of the images based on the fusion of features. To the best of our knowledge, this is the first work to explicitly evaluate the clustering effectiveness as a direct factor for diversification success instead of assessing final diverse rankings.

2. Related Work

Several approaches have been proposed to improve both relevance and diversity of image search results. In the image retrieval context, the work in [Sabetghadam et al. 2015] explored ensemble of clusters and fusion methods. Their diversification solution learns the best cluster set $C = (A, F, Di)$, where A is the clustering algorithm, F is the feature, and Di the distance measure. Therefore, they combine the results of distinct C by applying ensemble clustering based on the frequency that any two images end up in the same cluster. Besides that, they use two fusion methods, weighted linear and Bayesian inference. Hence, the relevance and diversity scores of each document were combined to generate final result. Although some stand out results in training phase, the fusion approaches did not overcome the ensemble of clusters.

The work in [Spyromitros-Xioufis et al. 2015] deals with the task using a rule-based fusion method. The authors proposed a task-specific supervised definition for relevance. Hence, instead of using the similarity to the query images to compute the relevance, they used the query images and the ground-truth data to train classifiers. A multimodal ensemble of classifiers was proposed to assign the relevance scores to images. The final result was obtained by jointly balancing relevance and diversity in a greedy reranking fashion.

Different from above, we evaluate several clustering algorithms and features. In order to boost the clustering effectiveness and maximize the information gain from diversity, we applied an ensemble clustering method to refine the results of the best combinations of the clustering algorithms and features. In [Sabetghadam et al. 2015], the authors also applied fusion methods using separate relevance and diversity measures. In our analysis, to achieve improvements in the clustering quality, we applied a feature fusion approach before constructing the clusters for diversification. Consequently, the selection and ranking step would be able to choose more diverse representative images. In turn, the work in [Spyromitros-Xioufis et al. 2015] achieved the best results using principal components analysis (PCA) on a single convolutional neural network feature. As an extension, our approach explored the dimensionality reduction on the fusion of the best performing individual features.

3. Proposed Analysis

The main goal of this work is attenuating the semantic gap in clustering results and consequently reduce the diversity gap. Therefore, we proposed applying features fusion using unsupervised learning techniques: ensemble clustering and principal component analysis (PCA). The ensemble clustering consists on combining multiple structures from different clusterings of the same data [Jain 2010]. PCA is a method to extract a lower-dimensional representation of the data [Han et al. 2012]. The objective of improving the clustering is to build more cohesive structures regarding the groups associated semantically.

In our experiments, the ensemble clustering was used for fusing multiple structures built by a clustering algorithm and multiple image features. Each instance of the clustering used only one feature. The clusterings were combined using the CSPA approach [Strehl and Ghosh 2002]. The result is a set of clusters from the consensus between the individual clusterings. We apply the same algorithm with distinct image representations based on a preliminary analysis using quality measures. In general, we observed close clustering effectiveness regardless the algorithm used. In turn, some features stand out on allowing the algorithms to achieve the highest effectiveness. Hence, for discussion, we explore the best performing features and a representative clustering algorithm.

For the PCA method, we concatenate the best performing features and select the most representative components from the resultant vector. A specific number of components was selected for each query in order to keep 90% of the variance. In our work, the objective was to find more effective image representations by creating better feature vectors with the fusion of features. We proposed to analyze the improvements of these approaches specifically on diversity promotion. Therefore, we evaluated the ensemble and PCA fusion approaches by measuring the quality of the clusters in an optimistic scenario containing only relevant images in the clustering task, which allows us to focus on the diversity gap. Therefore, we could properly assess the clustering quality and avoid noise interference since only relevant images are considered for final diversity assessment.

4. Experimental Setup

We evaluated the fusion approaches using the image collection from the *Retrieving Diverse Social Images Task* [Ionescu et al. 2015]. The experiments were performed with 153 queries, corresponding to 45,375 images from the development set. Each query has roughly 300 images obtained from Flickr. Besides the images, the collection provides textual metadata, and relevance and diversity ground-truth [Ionescu et al. 2015]. We selected 36 features [Calumby et al. 2015] for analysis, which are grouped based on the property encoded: Color (ACC, BIC, CM, CM3x3, CN, CN3x3, CLD, CSD, JCH, LUM, OPHIST, SCD, and SCH); Structure (CNN adapted, CNN generic, HOG, HSM, and WSA); Texture (CEDD, EHD, FCTH, Gabor, GIST, GLRLM, GLRLM3x3, JCD, LAS, LBP, LBP3x3, PHOG, and Tamura); Textual measures (Cosine, BM25, Dice, Jaccard, and TF-IDF, computed over the description, tags or title of the image from the metadata).

We evaluated six clustering algorithms. The partitional algorithm was k-Medoids since it has been used with success in recent works [Calumby et al. 2015]. The hierarchical algorithms evaluated were: single-link [Gan et al. 2007], complete-link [Gan et al. 2007], average-link [Gan et al. 2007], BIRCH [Zhang et al. 1996] and Chameleon [Karypis et al. 1999]. They were selected due to the intrinsic relationship with

the diversity task in terms of organizing the data into subtopics, and, besides that, promising results have been reported [Han et al. 2012]. In a preliminary evaluation, we defined the best performing features for each algorithm and the clustering results were assessed with extrinsic matching-based measures using human-generated diversity ground-truth (visual clusters). In particular, we computed the following measures: purity, maximum matching and F-measure [Zaki and Meira Jr 2014]. These measures presented highly correlated results. Hence, we present here only purity values. The purity measure quantifies to what degree a cluster c_i has objects from the ground-truth partitioning t_j, and is defined as $purity_i = \frac{1}{n_i} \max_{j=1}^{k} \{\Omega_{ij}\}$, where n is the number of objects in the clustering and Ω_{ij} is the number of common objects between a cluster and its most similar ground-truth partition. To assess how pure a clustering is, the sum over the number of common objects between clusters and ground-truth partitions is performed, which is averaged by the total number of objects in the clustering.

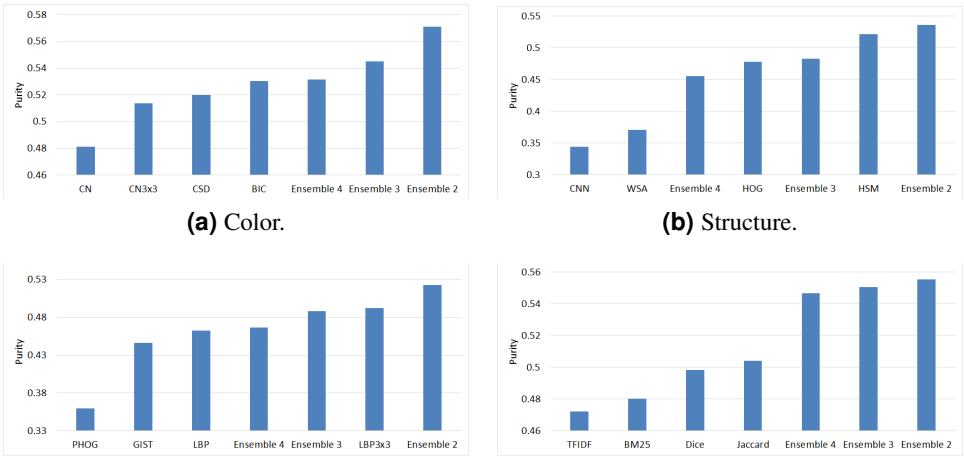
5. Results and Discussion

Our experiments were conducted to validate the proposed fusion approaches in terms of clustering effectiveness. Due to the general similar quality results obtained by the six clustering algorithms in preliminary experiments, we present the results only for the complete-link algorithm. For statistical significance definition, we applied the Wilcoxon's Signed-Rank Test with p-value < 0.05. In the results, the four features presented in each category are the best performing for the algorithm.

Figure 1 presents the sorted results of the complete-link instances with a single feature and for their ensemble. In Figure 1 the *Ensemble* methods represent the fusion of those features. Therefore, Ensemble 2, 3, and 4 correspond to the fusion using the 2, 3, 4 best performing features, respectively. For instance, in Figure 1a we present the results for the color-based descriptors: Ensemble 2 (BIC+CSD); Ensemble 3 (BIC+CSD+CN3x3); and Ensemble 4 (BIC+CSD+CN3x3+CN). In turn, the same pattern is followed for structure (Figure 1b), textual (Figure 1d), and texture (Figure 1c) features.

The results show that the ensemble of features provide a significant increase in purity. For all categories of features, the Ensemble 2 instance achieved the best results. Therefore, it indicates a great potential in combining algorithm instances where each instance look at the images using a distinct feature. Combining the top performing features led to a better final clustering. In general, all ensemble methods achieved statistically superior effectiveness in relation all 36 individual features. The only exceptions, considered as equivalent, were: for color, Ensemble 3 and BIC; for texture, Ensemble 2 and LBP3x3; and for structure, Ensemble 3 and HOG. The effectiveness decrease with the addition of more features may be a consequence of higher disagreement during the fusion step as a consequence of the semantic gap. In terms of efficiency, the cost of this method is quadratic, as a consequence of the computational and storage complexity of CSPA.

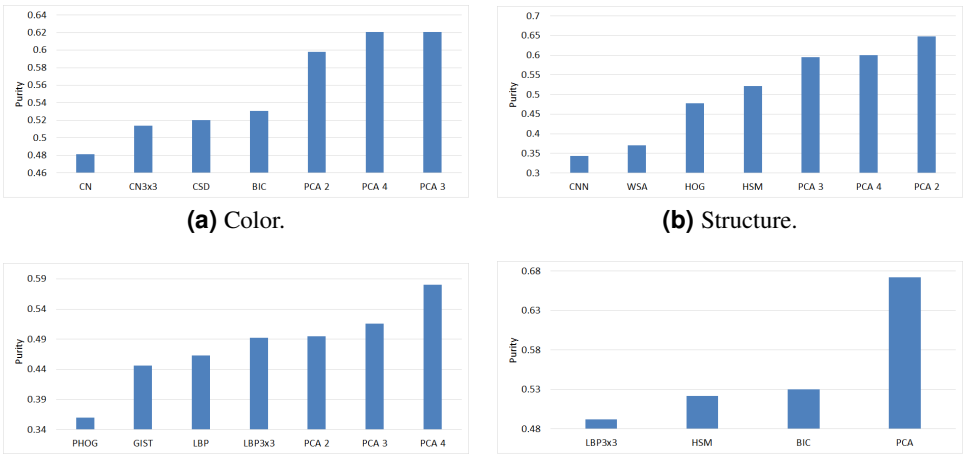
Figure 2 presents the purity effectiveness when the PCA was applied over the fusion of visual features. The top performing features were combined. The fusion pattern is equal to the one used for the ensemble. Hence, PCA 2 refers to the principal components from the two best performing features. The results demonstrate a stable gain using PCA. For all feature categories, all PCA fusions statistically outperformed all the individual features, except for PCA 2 and LBP3x3. Furthermore, in a per category comparative, this



(a) Color. **(b)** Structure.

(c) Texture. **(d)** Textual.

Figure 1. Ensemble clustering results sorted by purity.



(a) Color. **(b)** Structure.

(c) Texture. **(d)** Inter-category.

Figure 2. PCA-based results sorted by purity.

fusion approach also statistically outperformed the ensemble results. Figure 2d presents the PCA results for the combination of the best descriptor in each category. Finally, in terms of efficiency, the PCA, assuming it is an off-line procedure, may reduce the retrieval cost and demand less memory for indexing.

6. Conclusions

In the literature on diversity-based retrieval, the direct assessment of clustering quality is not discussed. We hypothesized that improving the quality of clusters, would allow selecting more representative images. Hence, we proposed to apply two fusion approaches aiming at improving the clustering for diversity while considering its intrinsic dependence on the performance of the low-level features. We have experimentally shown that it is possible to reduce the diversity semantic gap by jointly considering multiple features. The superiority of the fusion-based methods were demonstrated as statistically significant

when compared to individual features. These experimental results suggest that adequate features fusion is a promising alternative to reduce the semantic and diversity gaps, which may directly boost alternative diversity-oriented retrieval systems. As future work, we intend to conduct novel experiments considering a real scenario to assess the robustness of the approaches regarding noisy images. Besides that, the decreasing performance of ensemble clustering when more features are considered will be further investigated.

References

Boteanu, B., Mironica, I., and Ionescu, B. (2015). Hierarchical clustering pseudo-relevance feedback for social image search result diversification. In *CBMI*, pages 1–6.

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

Gan, G., Ma, C., and Wu, J. (2007). *Data clustering: theory, algorithms, and applications*, volume 20. Siam.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and techniques*. Morgan Kaufmann.

Ionescu, B., Popescu, A., Lupu, M., Gînsca, A.-L., and Müller, Henning, B. B. (2015). Retrieving diverse social images at mediaeval 2015: Challenge, dataset and evaluation. In *MediaEval*.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recogn. Lett.*, 31(8):651–666.

Karypis, G., Han, E.-H., and Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75.

Liang, S., Ren, Z., and De Rijke, M. (2014). Fusion helps diversification. In *ACM SIGIR*, pages 303–312. ACM.

Ounis, I., Macdonald, C., and Santos, R. L. (2015). Search result diversification. *Found Trends Inf Ret*, 9(1):1–90.

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

Spyromitros-Xioufis, E., Popescu, A., Papadopoulos, S., and Kompatsiaris, I. (2015). Usemp: Finding diverse images at mediaeval 2015. *Working Notes of MediaEval*.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *JMLR*, 3(Dec):583–617.

Veltkamp, R. C. and Tanase, M. (2002). *Content-Based Image and Video Retrieval*, chapter A Survey of Content-Based Image Retrieval Systems, pages 47–101.

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

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: an efficient data clustering method for very large databases. In *ACM SIGMOD*, volume 25, pages 103–114.