An Adaptive Clustering Approach for the Diversification of Image Retrieval Results

Maia Zaharieva[*]

Interactive Media Systems Group, Vienna University of Technology, Austria
Multimedia Information Systems Group, University of Vienna, Austria
maia.zaharieva@tuwien.ac.at

ABSTRACT

In this paper, we explore the application of an adaptive clustering approach for the diversification of image retrieval results in the context of the MediaEval 2016 Retrieving Diverse Social Images Task. The proposed approach exploits available textual descriptions, the visual content of the images, and a set of common clustering techniques to select the best combination for each image query individually and in an unsupervised manner.

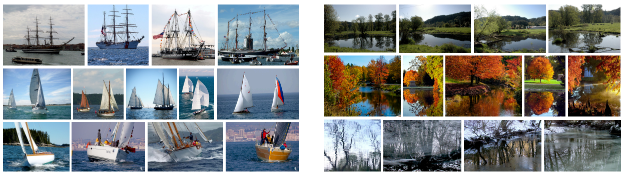
1. INTRODUCTION

The immense amount on publicly available media content commonly challenges end users in making use of the broad variety of accessible data. As a result, a lot of recent research focuses on the optimization of retrieval results in terms of improved relevance estimation and increased diversification [3, 6, 13, 21]. The MediaEval *Retrieving Diverse Social Images* task fosters the development and comparability of algorithms in this context [12]. The goal of the task in 2016 is to refine a set of images retrieved from Flickr as result of a general (and often multi-topic) query.

Increasing the relevance commonly leads to decreased diversity of the underling image set and vice versa. The magnitude of this reciprocal effect is difficult to estimate for different data settings. Therefore, in order to exploit the real potential of a clustering-based approach for data diversification, we consider all images as relevant for the given query. To address a general image retrieval scenario with arbitrary queries, we only consider commonly available textual information (title, description, tags) and the visual content of the images. The proposed approach does not make any assumptions about the initial image query or data characteristics. Moreover, the approach autonomously selects the best combination of image descriptions and clustering approach for each query individually and, thus, it is fully adaptive for different queries and data settings. Preliminary experiments demonstrate the generalization ability of the proposed approach and both its potentials and limitations.

2. APPROACH

The fundamental assumption behind the proposed approach is that different queries require for different features to efficiently describe the retrieval results in terms of diversification of the final image set. Figure 1 shows examples for desired image groupings for two queries: *sailing boats on sea* and *trees reflected in water*. While the results of the first query indicate potential relevance of the overall composition and edge-based descriptors, the second query suggests the use of color-based descriptors. On the contrary, color information is less meaningful for the first query since the retrieved images exhibit common color settings. Similarly, edges do not provide enough discriminative power to support the building of the desired image groupings for the second query. Therefore, in our study we consider a set of commonly employed visual- and text-based descriptors to represent the image data. Next to the two convolutional neural network (CNN)-based descriptors provided by the organizers [12], we additionally employ the first 36 coefficients of the discrete cosine transform (DCT) [1], intensity histogram (IH) [10], KANSEI shape descriptor [15], and six MPEG-7 visual descriptors [4, 18]: color layout (CL), color structure (CS), edge histogram (EH), homogeneous texture (HT), region-based shape (RS), and scalable color (SC). As text-based features we consider the well-established term frequency-inverse document frequency (TF-IDF) [14]. We compute the TF-IDF vector for each image using the available textual description (title, tags, and descriptions) in combination and individually. The textual descriptions are first preprocessed to increase their expressiveness, i.e., we remove potential occurrences of the corresponding user name, web links, and stopwords and we additionally stem all remaining terms.



(a) *sailing boats on sea* (b) *trees reflected in water*

Figure 1: Retrieval results for two image queries. Each line corresponds to a desired image groupings.

The unsupervised detection of existing groupings in a dataset is commonly performed by means of a clustering algorithm. However, the choice of a clustering approach is not a trivial decision. Different clustering approaches commonly address different data characteristics. Furthermore, potential clustering parameters usually require for an additional parameter tuning process. In this context, clustering internal validation indices are commonly applied in order to

[*] This work has been partly funded by the Vienna Science and Technology Fund (WWTF) through project ICT12-010.

MediaEval 2016 Workshop, October 20-21, 2016, Hilversum, Netherlands.

Table 1: Optimal solution: visual features. Numbers in the cells correspond to topic IDs.

Feature	AP	EM	KM	XM	$\sum$
CNN ad	58	11, 25	29	38, 51	6
CNN gen			4, 33, 39, 59	17, 54	6
DCT	48	19, 60, 67	50		5
IH	45	2	28, 53	6, 65	6
KANSEI shape		9, 20, 26, 36, 46	13, 57		7
MPEG7 CL		22, 47, 55	5, 31, 61	12, 18, 21	9
MPEG7 CS		15	7, 62, 70		4
MPEG7 EH		32, 49, 68	14, 34, 40, 42, 52		8
MPEG7 HT		1, 44	8, 27, 63	24	6
MPEG7 RS		23, 69	10, 41	35, 64, 66	7
MPEG7 SC		3, 43, 56		16, 30, 37	6
$\sum$	3	25	26	16	70

assess the quality of a clustering solution in terms of compactness and/or separability of the detected clusters [16]. We exploit the performance of several, broadly employed validation indices covering different aspects of a clustering solutions: 1) compactness in terms of sum of squares and C-index [11], 2) separability by means of single linkage distance between two clusters, 3) combination of compactness and separability in terms of Calinski-Harabasz [5], Davies-Bouldin [7], and Silhouette [20], and 4) consistency comparison measures by means of Gamma and Tau indices [2]. We employ the clustering validation measures to select the best combination of image feature and clustering algorithm for each query individually. As clustering methods we investigate two model-based approaches: Affinity Propagation (AP) [9] and expectation maximization (EM) [8], and two partitional approaches: k-means (KM) [17] and X-means (XM) [19]. The clustering methods were selected for their efficiency. Additionally, the only parameter tuning concerns the potential specification of the expected number of clusters, k. In this case, we perform clustering for various settings, $k = \{5, 10, 20, 30, 40, 50\}$, and consider each clustering solution individually. The final selection of a clustering solution for a given query is based on the quality assessment of all possible combinations between the considered clustering approaches and the employed image features. The final selection of images from the clusters follows a Round-Robin approach according to the Flickr-provided relevance scores. We start by selecting the image with the best relevance score from each cluster. These images, sorted in ascending order, constitute the m highest ranked results, where m is the number of detected clusters. The selected images are removed from their clusters and the selection process is repeated until the required number of retrieved results is achieved.

3. EXPERIMENTAL RESULTS

In our first experiment we investigate whether or not different datasets require for different features. For this purpose we perform clustering on the development dataset using all possible combinations between the considered clustering approaches (including potential clustering configurations) and the employed visual- and text-based features. The best clustering solution is selected using the ground truth information in terms of highest F1@20-score. Due to space limitations Table 1 shows an overview of the optimal clustering combinations for each topic (query) using the visual features only. The achieved results are presented in Table 2 (see *Optimal solution*). The balanced distribution of the clustering solutions across the employed visual features indicate that different image queries favor different features.

Table 2: Experiments on the development dataset.

Approach	P@20	CR@20	F1@20
Baseline: Flickr	0.6979	0.3717	0.4674
Optimal solution:			
visual features	0.8179	0.6575	0.7122
text features	0.8136	0.6453	0.7043
visual + text	0.8250	0.6634	0.7186
Best performing fixed settings:			
visual features	0.6657	0.4453	0.5237
text features	0.6636	0.4274	0.5045
visual + text	0.6657	0.4453	0.5237
Proposed adaptive approach:			
visual features	0.6500	0.4398	0.5123
text features	0.6729	0.4230	0.5029
visual+text	0.6286	0.4061	0.4803

Table 3: MediaEval 2016 benchmark results.

Run	Configuration	P@20	CR@20	F1@20
run 1	Adaptive, visual features	0.5141	0.4024	0.4292
run 2	Adaptive, text features	0.5406	0.4130	0.4463
run 3	Adaptive, visual+text features	0.5430	0.4130	0.4471
run 4	Fixed settings, visual features	0.4969	0.3603	0.4006

Additionally, the achieved performance in terms of F1@20-score significantly outperforms the baseline defined by the original Flickr result. The best performing clustering solution considers the DCT feature in combination with the KM clustering approach and the C validation index. This result is notably lower than the achievable performance. Additionally, the combination is selected based on experiments on a single dataset and, thus, overfitting the data. Experiments with the performance of the clustering validation indices indicated the C-index as the best performing validation measure stressing the importance of compactness for the final clustering solution. The results achieved using the proposed adaptive approach in combination with the C-index are presented in Table 2. The results are slightly lower than the best performing fixed setting yet not overfitting the data.

Table 3 summarizes the results on the test set. *Runs* 1 − 3 employ the adaptive approach while *run 4* exploits the best performing fixed settings on the development set as proof-of-concept for its overfitting. The proposed approach adapts well to the new dataset indicated by the difference in the selected features and clustering approaches. The fixed selection of image feature and clustering approach in *run 4* performs lowest as expected. Overall, the results are lower in comparison to the development dataset. However, this can only be evaluated in relation to the baseline which is not available for the test set yet.

4. CONCLUSION

In this paper we presented an initial study of the applicability of an unsupervised and adaptive clustering-based approach for the diversification of image retrieval results. The results indicate that different image queries favor different image representations and different clustering methods. The optimal solution for a dataset achieves an outstanding performance. However, the considered validation indices could not reflect the optimal solution. This might be due to the fact that different clustering solutions require for different validations. Our future work will exploit the potential of combining clustering validation indices in order to approach the best achievable performance.

5. REFERENCES

[1] N. Ahmed, T. Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974.

[2] F. B. Baker and L. J. Hubert. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70(349):31–38, 1975.

[3] G. Boato, D.-T. Dang-Nguyen, O. Muratov, N. Alajlan, and F. G. B. De Natale. Exploiting visual saliency for increasing diversity of image retrieval results. *Multimedia Tools and Applications*, 75(10):5581–5602, 2016.

[4] M. Bober. Mpeg-7 visual shape descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):716–719, Jun 2001.

[5] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

[6] D. T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F. G. B. D. Natale. A hybrid approach for retrieving diverse social images of landmarks. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, June 2015.

[7] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[9] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[10] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice-Hall, Inc., 3rd edition, 2006.

[11] L. J. Hubert and J. R. Levin. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 83(6):1072–1080, 1978.

[12] B. Ionescu, A. L. Ginsca, M. Zaharieva, B. Boteanu, M. Lupu, and H. Müller. Retrieving diverse social images at MediaEval 2016: Challenge, dataset and evaluation. In *MediaEval 2016 Multimedia Benchmark Workshop*, 2016.

[13] B. Ionescu, A. Popescu, A.-L. Radu, and H. Müller. Result diversification in social image retrieval: a benchmarking framework. *Multimedia Tools and Applications*, 75(2):1301–1331, 2016.

[14] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[15] H. Kobayashi, Y. Okouchi, and S. Ota. Image retrieval system using KANSEI features. In *Pacific Rim International Conference on Artificial Intelligence: Topics in Artificial Intelligence*, pages 626–635, 1998.

[16] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu. Understanding of internal clustering validation measures. In *IEEE International Conference on Data Mining (ICDM)*, pages 911–916, 2010.

[17] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

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

[19] D. Pelleg and A. W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *International Conference on Machine Learning (ICML)*, pages 727–734, 2000.

[20] P. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987.

[21] E. Spyromitros-Xioufis, S. Papadopoulos, A. L. Ginsca, A. Popescu, Y. Kompatsiaris, and I. Vlahavas. Improving diversity in image search via supervised relevance scoring. In *ACM International Conference on Multimedia Retrieval*, pages 323–330, 2015.