A Unified Framework for Retrieving Diverse Social Images

Maia Zaharieva[1,2] and Patrick Schwab[1]
[1]Multimedia Information Systems, Faculty of Computer Science, University of Vienna, Austria
[2]Interactive Media Systems, Institute of Software Technology and Interactive Systems,
Vienna University of Technology, Austria
maia.zaharieva@[univie|tuwien].ac.at, a0927193@unet.univie.ac.at

ABSTRACT

In this paper we explore the performance of a generic, unified framework for the retrieval of relevant and diverse images from social photo collections. The approach allows for the easy evaluation of different visual and textual image descriptions, clustering algorithms, and similarity metrics. Preliminary results show strong dependance between the choice of underlying technology and similarity metric, and the achieved performance.

1. INTRODUCTION

The immense daily growth of publicly shared media bears both potentials and challenges for automated media analysis and understanding. Currently, image access and retrieval is usually based on user-provided and camera-generated metadata. Although, more and more available, such metadata often suffers limitations such as imprecise capture time and GPS information or misleading and ambiguous textual descriptions. Additionally, the large number of shared items often results in a high-number of visually highly similar data. This challenge is the focus of the MediaEval 2014 *Retrieving Diverse Social Images Task* [2]. The aim of the task is the refinement of location images retrieved from Flickr while taking into consideration both their relevance and diversity.

Previous work in the context of this task shows a broad field of possible approaches ranging from re-ranking and clustering to greedy optimization and graph representations [3]. Several authors propose different systems for different feature types (e.g., [1][5]) that impedes the reasoning about the selection of an approach or particular features. Furthermore, some methods build upon assumptions that hold true in a limited setting only (e.g., relevance of an image is related to the number of views or the length of the descriptions [4]). While most of the presented approaches employ a combination of a re-ranking (for relevance improvement) and a clustering (for ensuring diversification) method, we build a unified framework that allows for a thorough evaluation of various textual and visual features, clustering algorithms, and similarity metrics.

2. APPROACH

We employ a multi-stage approach for the retrieval of diverse social images. The workflow passes three main stages: 1) relevance ranking of input images, 2) image clustering for diversification, and 3) final image selection. The initial set of input images may be optionally pre-processed in order to filter potentially irrelevant images, such as images with a human as main subject.

In the first stage, *relevance ranking*, each image of the input set is first represented by a feature vector $\vec{v}$, where $\vec{v}$ is a concatenation of the standardized z-scores of the feature descriptors $d_1, ..., d_n$:

$$\vec{v} = zscore(\big|d_1, \quad ..., \quad d_n\big|) \tag{1}$$

Since provided Wikipedia photos are per definition representative [2], we additionally compute a representative feature vector $\vec{v}_r$ for each referenced Wikipedia image. Following, the relevance score, s, of an image is defined as the smallest distance between its feature vector $\vec{v}$ and all $\vec{v}_r$ from the set of representative feature vectors W:

$$s = \min_{\vec{v}_r \in W} distance(\vec{v}, \vec{v}_r) \tag{2}$$

The aim of the second stage, *image clustering*, is to find groups of similar images that can be used to diversify the final retrieval results. Note that, distance measures and image features at this step are not necessarily the same ones employed for relevance ranking.

The third and last stage of the approach, *final image selection*, combines the results of the previous steps to retrieve images that are both relevant and diverse according to the initial image set. For this stage we use a Round-Robin algorithm. 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 corresponding clusters and the selection process is repeated until the required number of retrieved results is achieved.

In general, the clustering algorithm, the metric used to compare the feature vectors, and the underlying image features (for both image ranking and image clustering) are up to choice. In our experiments we tested different clustering algorithms: k-means, Adaptive Hierarchical Clustering (AHC), MeanShift, and Lingo, several comparison metrics: Euclidean, city-block, χ^2, cosine, correlation, Mahalanobis, Spearman, Hamming, and Jaccard, and all visual and textual features provided by the organizers [2]: term frequency - inverse document frequency (TF-IDF), Color Naming Histogram (CN), Histogram of Oriented Gradients (HOG), Color Moments (CM), Locally Binary Pattern (LBP), Statistics of Gray Level Run Length Matrix (GLRLM), and

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

Table 1: Best feature-metric combinations for AHC.

Relevance ranking			Image clustering		
1	CM3x3	Euclidean	1	CM	χ^2
2	TF-IDF	Spearman		CM3x3	Euclidean
	SIFT	Euclidean	2	HOG	cosine
3	CM	χ^2	3	GLRLM3x3	χ^2
	LBP	χ^2		LBP3x3	χ^2
	LBP3x3	χ^2	4	GLRLM	Euclidean
4	HOG	cosine		LBP	χ^2
	GLRLM	χ^2		SIFT	Euclidean
	GLRLM3x3	χ^2	5	CSD	cosine
5	CSD	cosine		CN	Euclidean
	CN	correlation		CN3x3	Euclidean
6	CN3x3	Euclidean	6	TF-IDF	Euclidean

the corresponding spatial pyramid representations (3x3) in addition to Bag-of-Visual Words (BoVW) of dense SIFT descriptors.

3. EXPERIMENTS AND RESULTS

In our first experiments we compared the performances of the different clustering algorithms. Results on the development data set showed that AHC significantly outperforms k-means, MeanShift and Lingo for all explored features (significance t-test, $p < 0.001$). Thus, we employed AHC in all follow up experiments.

We conducted a thorough evaluation of the performance of the employed features at the two main stages of our approach: relevance ranking and image clustering. Table 1 summarizes the results by means of ranked feature lists. The reported feature rankings and the selection of corresponding best performing distance measures are the product of significance t-tests with overall $p < 0.003$. While the Color Naming Histograms (CN and CN3x3) are usually outperformed by any other feature, the Color Moments (CM and CM3x3) show robust performance in both the ranking and the clustering tasks. In contrast to the ranking, which is clearly dominated by the performance of CM3x3, TF-IDF and SIFT, image clustering using AHC is more robust and the difference in the performance of global and local features decreases to a large extent.

Eventually, we submitted four runs for the final evaluation (see Table 2 for the configurations). Table 3 shows the results for the submitted runs for both development and test datasets. Best performances are achieved by the combination of textual and visual information (*run3*). However, in the context of the test dataset, the differences between the performances of the different runs vanish. Overall, clustering recall (CR) remains relatively low due to the large number of irrelevant images building noisy clusters. In general, the achieved results outline the limitations of the available textual (and visual) information in assessing image relevance. This is mainly due to the fact, that user-provided textual descriptions on social media sites often contain ambiguous or irrelevant information. A possible approach to improve the results may consider occasionally available GPS data and employ external resources as additional source for information.

Table 2: Official runs configurations (V: visual, T: textual descriptors employed).

	Relevance ranking	Image Clustering
run1 (V)	CM3x3	SIFT
run2 (T)	TF-IDF	TF-IDF
run3 (VT)	TF-IDF	CSD
run5 (V)	CM3x3	CSD

Table 3: Evaluation results.

	Development dataset			Test dataset		
	CR@20	P@20	F1@20	CR@20	P@20	F1@20
run1	0.4426	0.7600	0.5552	0.3901	0.6646	0.4863
run2	0.4132	0.7250	0.5188	0.3909	0.6809	0.4888
run3	0.4484	0.7567	**0.5559**	0.3982	0.6732	**0.4949**
run5	0.4369	0.7617	0.5499	0.3915	0.6752	0.4897

4. CONCLUSION

In this paper we presented a generic, unsupervised framework for the evaluation of various visual and textual features, similarity metrics, and clustering approaches for the retrieval of diverse social images. Performed experiments aim at the evaluation of the potentials and limitations of the provided visual and textual descriptions and, thus, we refrain from employing any assumptions or external sources of information. Although, there are significant differences in the performances of single features, the top performing features prove to be highly interchangeable. Achieved results indicate that - for the given datasets - the crucial part of the process is not so much the diversification but more the assessment of image relevance.

Acknowledgment

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

5. REFERENCES

[1] D. Corney, C. Martin, A. Göker, E. S. Xioufis, S. Papadopoulos, Y. Kompatsiaris, L. M. Aiello, and B. Thomee. Socialsensor: Finding diverse images at mediaeval 2013. In *MediaEval 2013 Workshop*, 2013.

[2] B. Ionescu, A. Popescu, M. Lupu, A. L. Gînscâ, and H. Müller. Retrieving diverse social images at mediaeval 2014: Challenge, dataset, and evaluation. In *MediaEval 2014 Workshop*, 2014.

[3] B. Ionescu, A. Popescu, H. Müller, M. Menéndez, and A.-L. Radu. Benchmarking Result Diversification in Social Image Retrieval. *IEEE International Conference on Image Processing*, 2014.

[4] N. Jain, J. Hare, S. Samangooei, J. Preston, J. Davies, D. Dupplaw, and P. H. Lewis. Experiments in diversifying flickr result sets. In *MediaEval 2013 Multimedia Benchmark Workshop*, 2013.

[5] B. Vandersmissen, A. Tomar, F. Godin, W. D. Neve, and R. V. de Walle. Ghent University-iMinds at MediaEval 2013 Diverse Images: Relevance-Based Hierarchical Clustering. In *MediaEval 2013 Workshop*, 2013.