LAPI @ Retrieving Diverse Social Images Task 2013: Qualitative Photo Retrieval using Multimedia Content

Anca-Livia Radu[1,2]*, Bogdan Boteanu[1], Oana Pleş[1], Bogdan Ionescu[1,2]
[1]The Image Processing and Analysis Laboratory, University "Politehnica" of Bucharest, Romania
[2]Department of Information Engineering and Computer Science, University of Trento, Italy
ancalivia.radu@unitn.it, bionescu@imag.pub.ro

ABSTRACT

In this paper we attempt to solve the Retrieving Diverse Social Images task by proposing an enhanced version of the method in [2] and studying the influence of its parameters in achieving high retrieval result diversification and relevance.

Keywords

Image search results diversification, visual and textual descriptors.

1. INTRODUCTION

The 2013 Retrieving Diverse Social Images Task [1] challenged participants to develop algorithms for selecting a small subset of representative and diverse images that correctly and completely summarize a query. Participants were provided with a development dataset containing 50 locations and a testing dataset containing 346 locations. The images for both data sets were retrieved from *Flickr* using the name of the location as query and also using the name of the location and the GPS coordinates [1]. We dealt with the task by developing a computer vision and linguistic processing algorithm that only employs visual and/or textual descriptors [2].

2. PREVIOUS WORK

Re-ranking techniques are the closest to our approach. Re-ranking attempts to re-order the initial retrieval results by taking advantage of the visual content and the additional information, such as textual data. Many approaches have been proposed in the literature, from methods that revaluate relational facts about the entities by estimating a model parameter, to methods proposing functions to optimize a diversity criterion or methods selecting representative images for a local group in the set that cover as many distinct groups as possible and that incorporate an arbitrary pre-specified ranking as prior knowledge [3] [4] [5].

3. OUR APPROACH

Our method, as presented in the sequel, selects from a given set of *N* retrieved images a small set of *F* images that are relevant and diverse representations of the query. First, it ranks the images in terms of representativeness using the similarity to the rest of the set. Then, all images are clustered and a small number of diverse images coming from different clusters are selected. Finally, a diversity rank is given by means of the dissimilarity to the rest of the selected images. A mediation between the two ranks guarantees the representativeness and diversity in images:

step 1: each image in the initial set is described using different combinations of descriptors. Further, in order to assess image similarity, we compute the Euclidean distance between the corresponding feature arrays and then construct a Synthetic Representative Image Feature (SRI) by averaging all distances.

step 2: a N-dimensional array is obtained by computing for each image the average of the Euclidean distances to the rest of the images. The value of *SRI* is subtracted from the new array which is further sorted in ascending order. The position of each value in the sorted array will be the new rank in terms of representativeness for the corresponding image.

step 3: all re-ranked images are clustered in *M* clusters using a k-means approach.

step 4: for each cluster a SRI_j value is computed and a new re-ranking is performed. From each cluster, a small equal number of best ranked images are selected to totally sum *F* best representative images.

step 5: another array is obtained by computing for all *F* images previously selected the average of the Euclidean distances to the rest $F-1$ images. The new array is sorted in descending order and the position of each value in the sorted array will be the new rank in terms of diversity for the corresponding image.

step 6: the average between the representativeness and diversity ranks is computed, resulting another array which is sorted in ascending order. Images are, thus, arranged and returned according to their final position in the sorted array.

4. EXPERIMENTAL RESULTS

The performance of our approach is influenced by a series of parameters: the descriptors and the number *M* of clusters to be built from all the images. We will first calibrate the method by experimenting on the development dataset using the provided visual and textual descriptors (i.e., color histograms, Histogram of Oriented Gradients, color moments, Locally Binary Patterns, MPEG-7 color structure descriptor, run-length matrix statistics and spatial pyramid representation of these descriptors, textual models [1][1]) . Then, we report the official results obtained on the testset.

4.1 Results on devset

For the development dataset of 50 locations several tests were performed by varying the parameters of the method as previously mentioned. Thus, different visual and textual descriptors combinations were tested while the number *M* of clusters was independently modified to 10 and 20. Figure1 presents the results obtained for the

*This research was partially supported by the CUbRIK project, grant agreement n287704, FP7/2007-2013.

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

[1] textual models are represented through histograms.

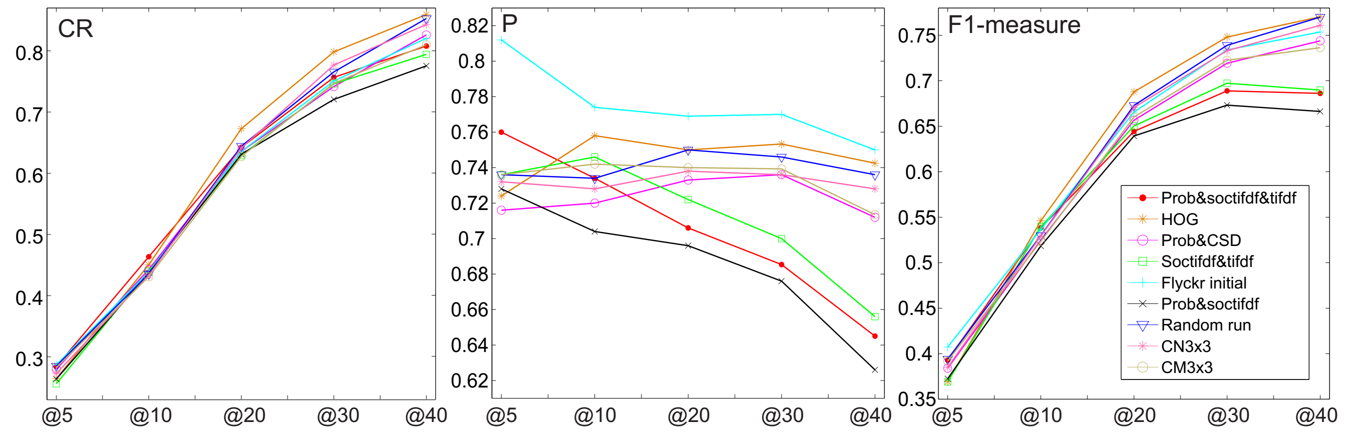

Figure 1: Official evaluation metrics for $M = 10$ (experiments on devset).

official evaluation metrics (cluster recall at 10 - $CR@10$, precision at 10 - $P@10$ and the harmonic mean of $CR@10$ and $P@10$ - $F1 - measure@10$) when M equals 10. For space reasons, we didn't graphically include the results obtained for M set to 20, since the overall results are less accurate. As Figure 1 a) depicts, the combination of all textual descriptors (the probabilistic model plus TF-IDF weighting and Social TF-IDF weighting) returns the best results among all visual and/or textual combinations in terms of the main evaluation metric ($CR@10$). That one is closely followed by the solely HOG visual descriptors and the combination between a textual (the probabilistic model) and a visual descriptor (CSD). The rest of descriptors' combination depicted in Figure 1 a) also return close results to the top 3 combinations.

On the other hand, when taking in consideration both $CR@10$ and $P@10$, thus evaluating $F1 - measure@10$, the results were also very close between the top combinations in terms of $CR@10$.

Textual descriptors perform better because they explicate better, when chosen carefully, the content and the details of the images than the visual descriptors that depict them in a simplified way.

4.2 Official runs

Following the previous experiments, we submitted four official runs computed as following: run1 - visual information only (using HOG descriptor), run2 - textual information only (using all provided textual descriptor, i.e., probabilistic model, term frequency-inverse document frequency (TF-IDF) weighting and social TF-IDF weighting), run3 - textual and visual fused information (using probabilistic model and CSD descriptors) and run5 - everything allowed (using CM3x3 descriptor).

Partial average results obtained in the official runs on the testing dataset are displayed in Table 1. The overall results obtained using the expert annotation are close to the ones obtained using the crowd annotation in terms of precision. Instead, the evaluation on the crowd generated ground truth lead to significantly higher cluster recall.

For the expert annotation, the best results in terms of the main evaluation metrics ($CR@10$) are achieved using a combination of all provided textual descriptors, thus only textual information. Moreover, the same combination offers the best results when considering both $CR@10$ and $P@10$, thus evaluating $F1 - measure@10$.

For the crowd annotation, the best results for $CR@10$ are achieved for the combination between a visual and a textual descriptor. In terms of $F1-measure@10$, the general run obtained using CM3x3 visual descriptor returned the best results.

Table 1: Official evaluation results ($*$ - evaluation on a selection of 50 locations)

	run	P		CR		$F1$	
		@10	@20	@10	@20	@10	@20
expert	*run1*	0,6901	0,6889	0,3631	0,5533	0,4582	0,5915
	run2	0,717	0,7111	0,3774	0,5734	0,4736	0,6078
	run3	0,6684	0,6813	0,3498	0,5444	0,438	0,5795
	run5	0,7371	0,7254	0,3742	0,5614	0,4726	0,6067
*crowd**	*run1*	0,6878	0,6898	0,7281	0,8594	0,6676	0,7393
	run2	0,7163	0,7255	0,7407	0,8583	0,6941	0,7641
	run3	0,6796	0,6929	0,7514	0,8653	0,6675	0,744
	run5	0,7143	0,7327	0,7322	0,8606	0,6942	0,77

5. CONCLUSIONS

We have presented a method for refining a set of noisy images retrieved from the web in terms of representativeness and diversity. Based on an extensive evaluation, our method proves to achieve great potential that overcome the initial retrieval using a broad range of visual and textual descriptors, leading to a precision up to 0.7371.

6. REFERENCES

[1] B. Ionescu, M. Menéndez, H. Müller, A. Popescu, "Retrieving Diverse Social Images at MediaEval 2013: Objectives, Dataset and Evaluation", MediaEval 2013 Workshop, October 18-19, Barcelona, Spain, 2013.

[2] A.-L. Radu, J. Stöttinger, B. Ionescu, M. Menéndez, F. Giunchiglia, "Representativeness and diversity in photos via crowd-sourced media analysis". AMR, 2012.

[3] B. Taneva, M. Kacimi, G. Weikum, "Gathering and ranking photos of named entities with high precision, high recall, and diversity". Int. Conf. on Web Search and Data, 2010.

[4] T. Deselaers, T. Gass, P. Dreuw, H. Ney, "Jointly optimising relevance and diversity in image retrieval". ACM Int. Conf. on Image and Video Retrieval, 2009.

[5] X. Zhu, A. Goldberg, J. V. Gael, D. Andrzejewski, "Improving Diversity in Ranking using Absorbing Random Walks". Int. Conf. HLT-NAACL, 2007.