PeRCeiVe Lab@UNICT at MediaEval 2014 Diverse Images: Random Forests for Diversity-based Clustering

Concetto Spampinato, Simone Palazzo
Department of Electrical, Electronics and Computer Engineering
University of Catania
Viale Andrea Doria, 6 - 95125, Catania, Italy
{cspampin, simone.palazzo}@dieei.unict.it

ABSTRACT

In this paper we describe the work done by the Pattern Recognition and Computer Vision Laboratory (PeRCeiVe Lab) of the University of Catania (Italy) for the MediaEval 2014 Retrieving Diverse Social Images Task. The main challenge consists of retrieving, for a given topic, a set of photos which are relevant to the topic but also showing diverse views of it. We submitted four runs exploiting a common feature-independent clustering strategy based on random-forests for diversifying Flickr result images while preserving relevance.

1. INTRODUCTION

The goal of the MediaEval 2014 Retrieving Diverse Social Images Task [1] is to refine a list of location photos, retrieved from Flickr through textual queries. Although photos retrieved in Flickr are often relevant, e.g., depict partially or entirely the target location, a significant number is either noisy or redundant. The objective is, therefore, to filter out such photos in order to obtain a exhaustive, and compact summary for the considered location.

Conversely to most of the existing methods, our approach first looks for diversity of the Flickr photos by performing a diversity-based clustering and then removes the irrelevant clusters by computing the similarities between the clustered photos and information available on Wikipedia. Final re-ranking is carried out by re-clustering the relevant images and computing a diversity score for each location photo.

2. METHOD

The strategy employed for dealing with the Retrieving Diverse Social Images Task of MediaEval 2014 relies on a common, feature-independent framework consisting of the following steps:

- **Diversity-based clustering**. The goal of this clustering step is to assess dissimilarity between samples (i.e., location photos described with either visual or text descriptors) of a given location and to group them according to such dissimilarity. To do that, we use Random forest predictors which allow us to define a dissimilarity measure between observations [3]. For each tree of the forest, if samples/observations i and j land in the same terminal node, their similarity score is increased by one. The similarities are finally normalized by the number of trees.

  The similarities between samples build up a matrix, SIM, which is symmetric, positive definite and with values in the unit interval $[0, 1]$. The dissimilarity matrix is defined as $DISSIM_{ij} = \sqrt{1 - SIM_{ij}}$, which is employed as input for partitioning around medoids clustering [2]. Each obtained cluster is likely to show a specific view of the considered location;

- **Cluster filtering by relevance**. Starting from the diversity clusters we then perform a cluster filtering by relevance. In particular, let SW be the average of maximum similarities between all the topic samples and the content available on Wikipedia computed as follows:

$$SW = \frac{1}{N}\sum_{i=1}^{N} \max_{j \in Wiki} SIM(i, j) \qquad (1)$$

  with N being the number of topic samples, and $Wiki$ the number of samples describing Wikipedia location content (e.g., in case of visual features, $Wiki$ is the number of images on Wikipedia for that location, while when employing text features $Wiki = 1$ since only sample describing the entire Wikipedia page text is considered).

  For each cluster C we carry out an unsupervised hierarchical tree clustering on samples' features, thus obtaining C' clusters. After that, we scale the similarities between the samples of cluster C and the Wikipedia content by C', i.e., $SIM'_C = SIM(i, j)_{i \in C, j \in Wiki} \cdot C'$. This gives more weight to the samples which resemble mostly the Wikipedia content and that, at the same time, are diverse from other samples. The relevance score RS_C of cluster C is, eventually, computed as mean of SIM'_C and the cluster is removed if $RS_C \geq k \cdot SW$;

- **Final ranking according to a diversity score**. The samples obtained at the previous step are again clustered using unsupervised hierarchical clustering and re-ranked according to a diversity score, computed for a sample j, by integrating: 1) number of samples in near clusters: if sample j is in cluster C_j we count how main samples are in C_j, C_{j-1} and C_{j+1} and its score is $s1_j = \frac{1}{N_{C_j} + N_{C_{j-1}} + N_{C_{j+1}}}$. This favors again diversity as samples with many items in near clusters are

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

Table 1: Training and official test results obtained by our method for each of the submitted runs.

	Development set			Test set (official)		
Run	**P@20**	**CR@20**	**F1@20**	**P@20**	**CR@20**	**F1@20**
run1 (visual)	86.67%	43.10%	56.87%	74.80%	38.74%	50.34%
run2 (text)	78.17%	44.02%	55.59%	75.53%	39.02%	50.63%
run3 (visual-text)	85.33%	43.61%	56.93%	72.40%	37.88%	49.03%
run5 (any resources)	84.14%	42.97%	56.14%	72.93%	37.31%	48.49%

strongly penalized; 2) photo id distance $s2_j$ from sample j to all other samples in the list: photos with close IDs are likely to refer to a similar view of a location; and 3) a random factor $s3_j$ to enable exploration of new solutions. The final ranking of sample j is given by multiplying the three above scores.

3. EXPERIMENTS

3.1 Setup and features

The algorithm described in Sect. 2 depends on the number of trees used for the diversity based clustering and on the k threshold for cluster filtering by relevance, which were set, respectively, to 50 and 3.5 for visual features and 1.0 for text features.

For runs using only visual features, we used the visual descriptors provided by the task's organizers for each photo (including the Wikipedia ones) normalized between 0 and 1 and concatenated into a single 945-dimensional vector.

Textual descriptors were computed as TF-IDF vectors from a vocabulary made up of all words from titles, descriptions and tags for photos in the development and test sets, plus words extracted from Wikipedia page (available as part of the data provided by the task's organizers). In order to reduce the original vocabulary size, being too large (more than 90,000 words), we removed: 1) words shorter than four characters; 2) words starting with digits; 3) words with low maximum TF-IDF values, thus resulting in a vocabulary size of 51,136 terms.

For both the visual-based and the text-based classifiers, the random forest was trained on 10 locations (the remaining ones were used for testing), randomly selected from the ones available in the training set. Increasing the number locations led to higher training times without an actual improvement in accuracy on the training set.

3.2 Results and discussion

We submitted four runs whose results are given in Table 1:

- Run 1 (visual information only): we employed the algorithm described in Sect. 2 on the visual features, filtering out images: 1) having people, detected by the face detector in [4], as subjects, and 2) whose Euclidean distance between the location's GPS coordinates (provided as part of each topic's description) and their GPS coordinates (when provided) was over 10;
- Run 2 (text information only): the same algorithm was employed on the reduced-vocabulary TF-IDF descriptors;
- Run 3 (visual-text fusion): the results presented in this run were obtained by combining those computed for run 1 and run 2, i.e., by multiplying the ranking scores of the two ranked lists for images appearing in both lists and completing the final list with images of the list of Run 1;
- Run 5 (any resources): same algorithm as in Run 1, without applying the face and GPS-based filtering.

It is clear that our attempt of making the training phase as independent as possible from the development set succeeded only partially, since the results on the test set are sensibly lower than those on the training set. Also, textual features performed surprisingly better than visual ones, which had obtained the highest accuracy by far on the development set. The significant performance difference, in terms of precision, of the visual runs (about 12%) on the test set and the development set can be due to a different relevance distribution of visual features, while it seems that text features keep the same distribution on the two sets.

4. CONCLUSIONS

In this paper we describe our random-forest based approach for tackling the MediaEval 2014 Retrieving Diverse Social Images Task. Our method, applied to text and visual features indifferently, leverages on a diversity-based clustering using Random Forests and on noisy cluster filtering to increase relevance. Final ranking is made in order to favor diversity with respect to relevance.

As future improvement, we mean to better exploit the amount of information provided for development: in particular, the diversity ground truth will be used to improve intra-cluster split in the diversity-based clustering, and user credibility information will be integrated into the relevance estimation model.

5. REFERENCES

[1] B. Ionescu, A. Popescu, M. Lupu, A. L. Gînsca, and H. Müller. Retrieving Diverse Social Images at MediaEval 2014: Challenge, Dataset and Evaluation. In *MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain*, 2014.

[2] H.-S. Park and C.-H. Jun. A simple and fast algorithm for k-medoids clustering. *Expert Syst. Appl.*, 36(2):3336–3341, Mar. 2009.

[3] T. Shi and S. Horvath. Unsupervised Learning With Random Forest Predictors. *Journal of Computational and Graphical Statistics*, 15(1):118 – 138, 2006.

[4] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57:137–154, 2001.