Multimodal-based Diversified Summarization in Social Image Retrieval

Duc-Tien Dang-Nguyen[1], Giulia Boato[1], Francesco G. B. De Natale[1], Luca Piras[2], Giorgio Giacinco[2], Franco Tuveri[3], Manuela Angioni[3]
[1] DISI - University of Trento, Italy
[2] DIEE - University of Cagliari, Italy
[3] Center for Advanced Studies, Research and Development in Sardina, Italy
dangnguyen@disi.unitn.it, boato@disi.unitn.it, denatale@ing.unitn.it, luca.piras@diee.unica.it, giacinto@diee.unica.it, tuveri@crs4.it, angioni@crs4.it

ABSTRACT

In this paper, we describe our approach and its results for the MediaEval 2015 Retrieving Diverse Social Images task. The main strength of the proposed approach is its flexibility that permits to filter out irrelevant images, and to obtain a reliable set of diverse and relevant images. This is done by first clustering similar images according to their textual descriptions and their visual content, and then extracting images from different clusters according to a measure of user's credibility. Experimental results shown that it is stable and has little fluctuation in both single-concept and multi-concept queries.

1. INTRODUCTION

In the MediaEval 2015 Retrieving Diverse Social Images task [4], participants are provided with sets of images retrieved from Flickr, where each set is related to a location. However, these sets are normally noisy and redundant, thus, the goal of this task is to refine the initial results by choosing a subset of images that are relevant to the queried location in different views, times, and other conditions.

We propose here an improved method based on our previous approaches in [1] and [2]. The basic idea is to filter out the non-relevant images at the beginning of the process according to the rules of the task. Then, exploit textual and visual features, as well as the *user credibility* information by a multi-modal retrieval framework to have a diversified summarization of the queried images.

2. METHODOLOGY

The proposed method comprises 3 steps (see Fig. 1):

Filtering: The goal of this step is to filter out outliers by removing images that are considered as non-relevant. We consider an image as non-relevant by defining the following rules: (i) it contains people as the main subject; (ii) it was shot far away from the queried location; (iii) it received very few number of views on Flickr; and (iv) it is out-of-focus or blurred. Condition (i) can be detected by the proportion of the human face size with respect to the size of the image. In our method, the Luxand FaceSDK (luxand.com) is used as a face detector. Conditions (ii) and (iii) can be computed exploiting the provided user credibility information. In order to detect blurred images (rule iv), we estimate the focus by computing the sum of wavelet coefficients and decide if it is out-of-focus following the method in [3]. After this step, all the images left are considered as relevant and are passed to the next step.

Clustering: we propose to cluster similar images by constructing a particular clustering feature tree (CF tree) which is built based on the combination of textual and visual information. To this end, we exploit the characteristic of the BIRCH algorithm [6] to perform clustering in two main phases, namely the Global Clustering phase, and the Refining phase. While these two phases are intended to produce a high quality clustering results by using the same set of features, we used textual features to perform the first phase and we refined the clusters by using visual features instead. We computed a different set of textual features by performing the analysis of the provided textual data in order to reduce the noise of not relevant words. After this step, all images that are visually similar and have the same context (i.e., the textual information) are grouped into the same branch of the tree.

Summarization: Starting from the CF tree, the clusters can be obtained by applying the agglomerative hierarchical clustering algorithm on CF leaves to form the set of clusters. To choose the best images for summarizing the landmark, first the clusters are sorted based on the number of images, i.e., clusters containing more images are ranked higher. Then, we extract images from each cluster till the maximum number of required images is reached (e.g., 20 images). In each cluster, the image uploaded by the user with highest visual score is selected as the first image. If there is more than one image from that user, the image closest to the centroid is selected. If more than one image have to be extracted from a cluster to reach the exact number of images required to build the visual summary, we select the second image as the one which has the largest distance from the first image, the third image as the one with the largest distance to both the first two images, and so on.

3. RUN DESCRIPTION

We ran our model on the development set (devset, containing 153 location queries from 45.375 Flickr photos). According to the results, we choose the best features and the tuned parameters for each run and applied to the test set (containing 69 single-concept queries and 70 multi-concept

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

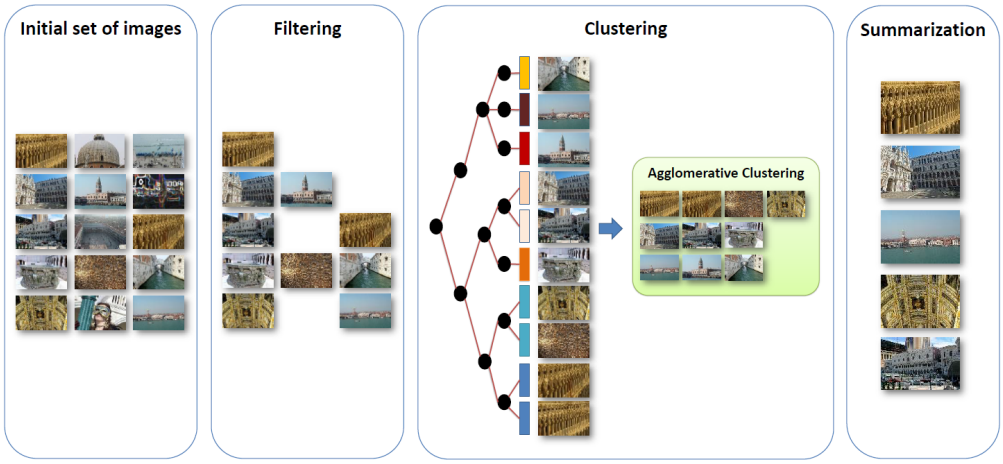

Figure 1: Schema of the proposed method

Table 1: Run performances on MediaEval Retrieving Diverse Social Images Task 2015 Test Set.

	Single Concept			Multi Concept			Overall		
Run	**P@20**	**CR@20**	**F1@20**	**P@20**	**CR@20**	**F1@20**	**P@20**	**CR@20**	**F1@20**
Run 1	0.6601	0.406	0.4902	0.6643	0.4226	0.5017	0.6622	0.4143	0.4959
Run 2	0.5993	0.37	0.4461	0.6636	0.4275	0.5028	0.6317	0.399	0.4747
Run 3	0.6181	0.3725	0.4538	0.67	0.4315	0.5088	0.6442	0.4022	0.4815
Run 4	0.6768	0.4131	0.5009	0.6921	0.4198	0.5052	0.6845	0.4165	0.5031
Run 5	**0.7362**	**0.4288**	**0.529**	**0.7607**	**0.4753**	**0.567**	**0.7486**	**0.4522**	**0.5481**

queries from 41.394 Flickr images) as follows:

Run 1: Color naming (CNM), color descriptor (GCD), histogram of oriented gradients (HOG) and local binary pattern (GLBP) are used. In the Summarization step, since we do not have the user credibility information in this run, the centroid of each cluster is selected as the first image.

Run 2: In this run, we refined text features by normalizing the text terms and removing stop-words, html tags and special characters from the given TF-IDF. Cosine similarity was used as the distance metric. The parameters are chosen similar to Run 1.

Run 3: The proposed method is applied on the combined features from run 1 and run 2 where TF-IDF is used first, then the visual features with Euclidean distance are applied after.

Run 4: In this run, we clustered the images by user. The order of the clusters is ranked based on the visual score (i.e., the cluster belong to the user with highest visual score will be selected first), then by face proportion, and so on with all the user credibility information. For each cluster, images are selected based on the number of views, i.e., the image with highest number of views is selected as the first image.

Run 5: In the first four runs, we applied the same method on both single-concept and multi-concept queries. However, in this run, we used two different methods for these two different cases. In the filtering step for single-concept queries, outliers are detected as follows: rule (i): the face size is bigger than 10% with respect to the size of the image, (ii) images that were shot farther than 15kms, (iii) images that have less than 25 views, and (iv) images that have f-score (focus measure) smaller than 20. For the multi-concept queries, only rule (iii) and (iv) were applied since there are many queries require images belong to multiple locations. We also removed images whose title and descriptors do not contain any word from the query. In the Clustering step, a similar clustering as Run 3 is applied for both types of query with the extra visual features: Dense SIFT and HOG2x2, extracted as the study in [5]. Text features were refined as described in Run 2. Finally, in the Summarization step, the same method as described in Section 2 were applied.

Table 2: Run performances on Development set.

Runs	P@20	CR@20	F1@20
Run 1	0.7268	0.4125	0.5188
Run 2	0.7229	0.4245	0.5127
Run 3	0.8000	0.4013	0.5266
Run 4	0.7012	0.4198	0.5015
Run 5	0.8517	0.4829	0.6102

4. RESULTS AND CONCLUSION

With the mentioned selected features and parameters, we obtained the highest $F1@20$, the official metric of the task, at Run 5 on both development and test sets with the values of 0.61 and 0.55, respectively. These results confirmed that removing outliers and combining textual, visual and user credibility information as run 5 significantly improved the performance with respect to the other runs (see in Table 1 and Table 2 the results on the test set and development set, respectively).

According to the results on the test set, we can state that the performances is stable and has little fluctuation in both single-concept ($F1@20 = 0.529$) and multi-concept ($F1@20 = 0.567$) queries.

5. REFERENCES

[1] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F. G. B. De Natale. Retrieval of diverse images by pre-filtering and hierarchical clustering. In *MediaEval*, 2014.

[2] 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*, 2015.

[3] J.-T. Huang, C.-H. Shen, S.-M. Phoong, and H. Chen. Robust measure of image focus in the wavelet domain. In *Intelligent Signal Processing and Communication Systems*, pages 157–160, Dec 2005.

[4] B. Ionescu, A. L. Gînscă, B. Boteanu, A. Popescu, M. Lupu, and H. Muller. Retrieving Diverse Social Images at MediaEval 2015: Challenge, Dataset and Evaluation. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, September 2015.

[5] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.

[6] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: An Efficient Data Clustering Method for Very Large Databases. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pages 103–114, 1996.