# Retrieval of Diverse Images by Pre-filtering and Hierarchical Clustering

D.-T. Dang-Nguyen, L. Piras, G. Giacinto
DIEE - University of Cagliari
Piazza D'armi, 09123 Cagliari, Italy
{ductien.dangnguyen, luca.piras, giacinto}@diee.unica.it

G. Boato, F. G. B. De Natale
DISI - University of Trento
Via Sommarive, 9 I-38123 Povo, Trento, Italy
boato@disi.unitn.it
denatale@ing.unitn.it

## ABSTRACT

In this paper, we describe our approach and its results for MediaEval 2014 Retrieving Diverse Social Images Task. The basic idea of our proposed method is to filter out non-relevant images at the beginning of the process and then construct a hierarchical tree which allows to cluster the images with different criteria on visual and textual features. Experimental results shown that it is stable and has little fluctuation with the number of topics.

## 1. INTRODUCTION

In MediaEval 2014 Retrieving Diverse Social Images task [2], 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 but reporting different views of the location, various perspective, different daytimes (e.g., night and day), etc.

The basic idea of the proposed method is to filter out the non-relevant images at the beginning based on the rules of the task and then use for clustering the BIRCH algorithm [4], that builds a hierarchical structure where nodes are the images, and edges represent the similarity between the linked nodes. This structure allows creating different clusters, according to the criteria used, and can also be used to remove outliers, i.e., non-relevant images that were not filtered out during the first step.

## 2. METHODOLOGY

The proposed method contains 4 steps (see Fig. 1):

- Step 1. Pre-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 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, Luxand FaceSDK [1] 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 [1]. After this step, all the images left are considered as relevant and are passed to the next step;

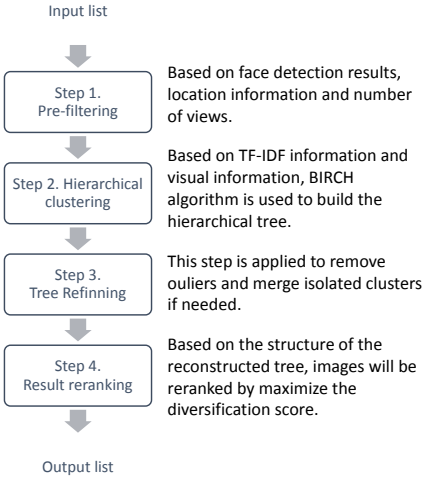


**Figure 1: Schema of the proposed method**

- Step 2. Hierarchical Clustering: In this step, we use the BIRCH clustering algorithm [4] on the provided visual and textual features. BIRCH allows to obtain an initial clustering result in large datasets with very low computational costs. Images that are similar to each other based on global visual features and textual information after this step are grouped into the same cluster or the same branch of the hierarchical tree;

- Step 3. Tree Refining: thanks to the initial tree constructed in the previous step, isolated clusters can be easily removed or merged to other branches by updating the tree, without modifying the clusters;

- Step 4. Result Re-ranking: the clusters are sorted based on the number of images, i.e., clusters contain more images are ranked higher. In each cluster, the image uploaded by the user who has highest visual score is selected as the first image. If there are more than one image from that user, the image closest to the centroid is selected. The second image is the one which

[1] http://luxand.com/


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

**Table 1: List of features used in the submitted runs.**

| | Visual features | Text features | User credibility | Other features |
|---|---|---|---|---|
| Run 1 | CNM, GCD, HOG, GLBP | - | - | - |
| Run 2 | - | TF-IDF | - | - |
| Run 3 | CNM, GCD, HOG, GLBP | TF-IDF | - | - |
| Run 4 | - | - | All credibility information | - |
| Run 5 | CNM, GCD, GLBP | TF-IDF | views, visual score | HOG2x2, f-Score, face size, GPS |

**Table 2: Run performances on Development set.**

| Runs | P@20 | CR@20 | F1@20 |
|---|---|---|---|
| Run 1 | 0.7783 | 0.4441 | 0.5592 |
| Run 2 | 0.7017 | 0.4245 | 0.5215 |
| Run 3 | 0.8000 | 0.4013 | 0.5266 |
| Run 4 | 0.6933 | 0.4116 | 0.5084 |
| Run 5 | **0.8367** | **0.4488** | **0.5737** |

**Table 3: Run performances on Test set.**

| Runs | P@20 | CR@20 | F1@20 |
|---|---|---|---|
| Run 1 | 0.7561 | 0.4439 | 0.5510 |
| Run 2 | 0.7232 | 0.4247 | 0.5289 |
| Run 3 | 0.7179 | 0.4191 | 0.5233 |
| Run 4 | 0.7175 | 0.4238 | 0.5252 |
| Run 5 | **0.8512** | **0.4692** | **0.5971** |

has the largest distance to the first image. The third image is chosen as the image with the largest distance to both the first 2 images, and so on.

Several similarities and metrics have been used: for the provided visual information, we use Euclidean distance, while with textual information, we use cosine similarity. About geo-tagged images, Haversine formula is used to compute the geographical distance between two locations.

# 3. RESULTS AND DISCUSSION

In order to find the best combination of features and parameters (the number of clusters, the inner parameters of BIRCH, and the thresholds to determine the outliers), we ran our model for all the provided features together with our own features. According to the results, we choose the best features and parameters for each run and applied to the test set as follows:

- Run 1 (*Visual*): Color naming (CNM), color descriptor (GCD), histogram of oriented gradients (HOG) and local binary pattern (GLBP) are used. In Step 4, since we cannot exploit user credibility information, the centroid of each cluster is selected as the first image.

- Run 2 (*Text*): The parameters are chosen similar to Run 1, but we used only TF-IDF information and measure the distances by cosine similarity.

- Run 3 (*Visual+Text*): The 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 (*User credibility*): Please notice that this run is allowed to use only the user credibility information, thus the proposed method is not applied. 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 (*All features*): All steps in the proposed method are applied in this run. In Step 1, outliers are detected as follows: (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. In Step 2, a similar clustering as Run 3 is applied. About the visual features, we replace the provided HOG features by HOG2x2 as presented in [3].

Table 1 summarizes all the features that have been used in our runs. With the mentioned selected features and parameters, we obtained the highest F1@20, the official metrics of the task. In particular we obtain the best results at Run 5 on both development and test sets with F1@20 values of 0.57 and 0.6, P@20 values of 0.84 and 0.85, and CR@20 values of 0.45 and 0.47, respectively. All scores on development set are reported in Table 2 while Table 3 shows the results on the test set for all runs. It can be noticed that visual features are crucial for achieving good performances (second best result). According to the results, we can state that the performances on test set and development set are consistent, proving that the proposed method is stable and has little fluctuation with the number of topics.

# 4. REFERENCES


[1] 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.

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

[3] 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.

[4] 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.