MUCKE Participation at Retrieving Diverse Social Images Task of MediaEval 2013

Anil Armagan Bilkent University, Department of Computer Engineering, 06800, Ankara, Turkey. anil.armagan@bilkent.edu.tr Adrian Popescu CEA, LIST, Vision & Content Engineering Laboratory, 91190 Gif-sur-Yvette, France. adrian.popescu@cea.fr Pinar Duygulu Bilkent University, Department of Computer Engineering, 06800, Ankara, Turkey. duygulu@cs.bilkent.edu.tr

ABSTRACT

The Mediaeval 2013 Retrieving Diverse Social Image Task addresses the challenge of improving both relevance and diversity of photos in a retrieval task on Flickr. We propose a clustering based technique that exploits both textual and visual information. We introduce a k-Nearest Neighbor (k-NN) inspired re-ranking algorithm that is applied before clustering to clean the dataset. After the clustering step, we exploit social cues to rank clusters by social relevance. From those ranked clusters images are retrieved according to their distance to cluster centroids.

1. INTRODUCTION

Photo sharing is a frequent activity on social media platforms, such as Flickr and Facebook. Since there is no quality control of the photos and of their annotations, retrieved images for a given query are usually not sufficiently relevant. Even when the relevance criteria is met, it is still difficult to obtain diversified representations [4].

Considering these challenges, the MediaEval 2013 Retrieving Diverse Social Images Task addresses the problem of diversification of social image search results (see [2] for the detailed description of the task and the dataset).

The dataset used in the task is collected from Flickr. Annotations are usually limited to a few number of tags and therefore incomplete [3]. Flickr images also come with social cues, such as user ID.

In the following, we summarize our efforts in solving the challenges mentioned above through exploiting both textual and visual features, as well as social cues.

2. FEATURES

2.1 Visual Features

We exploit visual features in order to overcome the sparsity of textual annotations since there are only few tags for each image. We use the Histogram of Gradients (HOG) in our experiments. In addition, we extracted GIST and bags of visual words (BOVW) based on dense SIFTs that proved to be efficient in large scale image retrieval. Dense SIFT descriptors are extracted using a codebook of size 1024. A spatial pyramid model with 2 levels is used and the resulting feature size is 8192. HOG, GIST and BOVW features capture different low-level characteristics of images and they can be combined to have more comprehensive visual representations. Since these features can be combined in different kind of runs, all features were L1-normalized in order for each of feature to have the same contribution, regardless of their size.

2.2 Textual Features

We exploit a classical TF-IDF weighting scheme to model textual information associated to points of interests (POIs). Different re-weighting schemes were tested and the best results were obtained when we took the square root of TF-IDF scores. The dimension of the model is equal to the number of unique tags associated to each POI. Given the limited number of images per POI (up to 150), the dimension is usually in the range of hundreds. Similar to visual features, L1-normalization is applied to textual features.

3. RESULT RERANKING

The initial result set of Flickr is noisy and we introduce a k-NN inspired approach that exploits visual and social cues to rerank results to reduce this noise. We considered all the images of the POI as a positive set and constructed a negative set of the same size by sampling images of other POIs from the collection. Then we compared the GIST features of each image to all other images' features from positive and negative sets. The resulting top five most visually similar images are retained to be considered in the next steps.

The top five neighbors from both the positive and the negative sets of each image that depict a POI are considered to produce a reranked results list according to social cues. For each image we counted the number of different users that contributed to the top five neighbors and the number of top five neighbors that belong to the positive set. Finally the average distance to the first five positive neighbors of the target image is considered. These criteria were cascaded to break ties. Images are then ranked by cascading the three scores described and we used the top 70%, 80% and 90% of the reranked images. These images are given as an input for the clustering process. The best results are obtained on the devset with 70% of the initial list retained and this threshold is retained for clustering.

4. CLUSTERING

k-means++ [1] algorithm is used to cluster the images of a topic using previously mentioned feature types for dif-

Copyright is held by the author/owner(s).

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

 Table 1: Run performances with three official metrics using expert ground truth

Run name	CR@10	P@10	F1@10
Run1 - visual	0.3892	0.7243	0.4905
Run2 - textual	0.3869	0.7333	0.489
Run3 - textual-visual	0.3848	0.7272	0.4868
Run5 - everything allowed	0.3742	0.7161	0.4753

ferent kind of runs. Different numbers for k value are used in experiments such as 10, 15 and 20. K value is selected as 15 since this value gives us the best results on evaluation metrics precision, recall and f1 measure at 10. Also k = 15is close to the official metrics (CR@10, P@10, F1@10) used in the evaluation.

5. CLUSTER AND RESULT RANKING

Clusters are not all born equal and we need to be able to rank them by probability of relevance of contained images. Inspired by [4], we exploit social cues for cluster ranking and propose a simple scheme that is based on user and date information. For each cluster, we count the number of different users that contributed to that cluster and the number of different dates when photos of that cluster were taken. The first count aims to prioritize clusters that are socially diverse while the second count aims to surface clusters that are temporally stable. Then we calculate the product of these two counts and consider it as a social ranking score. To break ties, we also use the number of images present in each cluster.

For each POI, we retain only the top 10 clusters obtained with the cluster ranking procedure and then diversify images by choosing one image from each cluster by descending similarity to the cluster centroid.

6. RESULTS AND DISCUSSION

To address the diversified social image retrieval problem, participants are asked to submit different types of runs. We submitted four runs that were produced by using different types of features and their combinations on the same dataset. The submitted four runs are described below: Run1 - visual is based on visual features. We concatenate HOG and GIST features described in Section 2.1. Run2 - textual is produced using only textual features described in Section 2.2. Run3 - textual-visual is produced using a combination of textual features described in Section 2.2 and GIST features in Section 2.1. Visual and textual features are concatenated and to produce feature vectors. Linear weighting is used with 0.7 and 0.3 weights that are given to visual and textual features respectively. These weights were empirically chosen by testing different combinations on the devset. Run5 - everything allowed is similar to Run 3 - textualvisual, the only modification being the replacement of HOG features by BOVW features.

The results in Table 1 are based on expert evaluation on 346 testset locations. Table 2 shows the average results of crowd sourcing evaluation carried out only on a subset of 50 locations from the testset via the CrowdFlower platform. Relevance ground truth is based on a majority voting scheme on the annotations and the diversity ground truth is determined by the same three annotators. Both tables show that there are only small differences between the four runs. We

Table 2	2: Run	performa	ances	\mathbf{with}	\mathbf{three}	official	met-
rics usi	ing cro	wd sourci	ng gr	ound	truth		

Run name	CR@10	P@10	F1@10
Run1 - visual	0.7446	0.7449	0.7135
Run2 - textual	0.7406	0.7204	0.6997
Run3 - textual-visual	0.7503	0.7245	0.705
Run5 - everything allowed	0.7332	0.7184	0.6951

were surprised to see how well the textual run (Run2) performed compared to visual and multimedia runs since we had expected visual diversification to work better for POIs, which usually have a limited number of visual aspects. The performance drop from Run3 to Run5, due to the replacement of HOG features with BOVW features also came as a surprise since the latter usually work well for retrieval processes over visually diversified datasets.

7. CONCLUSION

In this paper we propose a clustering based technique using textual and visual features. To re-rank the images before clustering we introduced a kNN inspired technique. This technique helped us to spot strongly connected images of a topic. We further apply cluster ranking method based on social cues to increase the diversity and the relevancy. Table 1 and Table 2 show that there are no big differences between different runs. Performance of Run2 - textual and Run5 everything allowed is a surprise for us. Since BOVW representation generally work well for retrieval processes. We would expect from the visual representations to improve the results more than the textual features.

8. ACKNOWLEDGMENT

This research was supported by the MUCKE project funded within the FP7 CHIST-ERA scheme and also Scientific and Technical Research Council of Turkey (TUBITAK) under grant number 112E174.

9. **REFERENCES**

- D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.
- B. Ionescu, M. Menendez, H. Muller, and A. Popescu. Retrieving diverse social images at mediaeval 2013: Objectives, dataset and evaluation. In *MediaEval 2013* Workshop, CEUR-WS.org, ISSN: 1613-0073, Barcelona, Spain, October 18-19 2013.
- [3] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev. To search or to label?: predicting the performance of search-based automatic image classifiers. In *Proceedings* of the 8th ACM international workshop on Multimedia information retrieval, MIR '06, pages 249–258, New York, NY, USA, 2006. ACM.
- [4] L. S. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 297–306, New York, NY, USA, 2008. ACM.