Retrieving Diverse Social Images at MediaEval 2013: Objectives, Dataset and Evaluation

Bogdan Ionescu
LAPI, University Politehnica of Bucharest, Romania
bionescu@alpha.imag.pub.ro

María Menéndez
DISI, University of Trento, Italy
menendez@unitn.it

Henning Müller
HES-SO, Sierre, Switzerland
henning.mueller@hevs.ch

Adrian Popescu
CEA-LIST, France
adrian.popescu@cea.fr

ABSTRACT

This paper provides an overview of the Retrieving Diverse Social Images task that is organized as part of the MediaEval 2013 Benchmarking Initiative for Multimedia Evaluation. The task addresses the problem of result diversification in the context of social photo retrieval. We present the task challenges, the proposed data set and ground truth, the required participant runs and the evaluation metrics.

1. INTRODUCTION

The MediaEval 2013 Retrieving Diverse Social Images Task addresses the problem of result diversification in the context of social photo retrieval. Existing retrieval technology focuses almost exclusively on the accuracy of the results that often provides the user with near replicas of the query. However, users would expect to retrieve not only representative photos but also diverse results depicting the query in a comprehensive and complete manner. Another equally important aspect is that retrieval should focus on summarizing the query with a small set of images, since most of the users commonly browse only the top retrieval results.

The task aims to foster new research in this area [1, 2] by creating a multi-modal evaluation framework specifically designed to encourage the creation of new solutions from various research areas, such as: machine analysis, human-based approaches (e.g., crowd-sourcing) and hybrid machine-human approaches (e.g., relevance feedback). Compared to other existing tasks addressing diversity, e.g., ImageCLEF 2009 Photo Retrieval [3], the main novelty of this task is in addressing the social dimension that is reflected both in its nature (variable quality of photos and of metadata) and in the methods devised to retrieve it.

2. TASK DESCRIPTION

The task is build around a tourist use case where a person tries to find more information about a place she is potentially visiting. The person has only a vague idea about the location, knowing the name of the place. She uses the name to learn additional facts about the place from the Internet, for instance by visiting a Wikipedia[1] page, e.g., getting a photo, the geographical position of the place and basic descriptions. Before deciding whether this location suits her needs, the person is interested in getting a more complete visual description of the place.

In this task, participants receive a list of photos for a certain location retrieved from Flickr[2] and ranked with Flickr's default "relevance" algorithm. These results are typically noisy and redundant. The requirements of the task are to refine these results by providing a ranked list of up to 50 photos that are considered to be both *relevant* and *diverse* representations of the query according to the definitions:

Relevance: a photo is *relevant* for the location if it is a common visual representation of the location, e.g., different views at different times of the day/year and under varying weather conditions, inside views, close-ups on architectural details, drawings, sketches, creative views, etc, which contain partially or entirely the target location. Photos of poor quality (e.g., severely blurred, out of focus, etc) as well as photos showing people in focus (e.g., a big picture of me in front of the monument) are not considered relevant.

Diversity: a set of photos is considered to be *diverse* if it depicts different visual characteristics of the target location, e.g., different views at different times of the day/year and under varying weather conditions, inside views, close-ups on architectural details, creative views, etc, with a certain degree of complementarity, i.e., most of the perceived visual information is different from one photo to another.

3. DATASET

The 2013 data set consists of 396 locations, spread over 34 countries around the world, ranging from very famous ones (e.g., "Eiffel Tower") to lesser known monuments (e.g., "Palazzo delle Albere"). They are divided into a development set containing 50 locations (*devset* - to be used for designing and validating the proposed approaches) and a test set containing 346 locations (*testset* - to be used for the official evaluation). Each of the two data sets contains data that was retrieved from Flickr using the name of the location as query (*keywords*), as well as using the name of the location together with its GPS coordinates (*keywordsGPS*).

For each location, the following information is provided: the name of the location, its GPS coordinates, a link to a Wikipedia description webpage, a representative photo from Wikipedia, a ranked list of photos retrieved from Flickr (up

[1] http://en.wikipedia.org/

[2] http://www.flickr.com/

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

to 150 photos per location; *devset* contains 5,118 images while *testset* 38,300 images)[3], an xml file containing metadata from Flickr for all the retrieved photos (i.e., photo title, photo description, photo id, tags, Creative Common license type, number of posted comments, the url link of the photo location from Flickr, the photo owner's name and the number of times the photo has been displayed), a set of global visual descriptors automatically extracted from the photos (i.e., color histograms, histogram of oriented gradients, color moments, local binary patterns, MPEG-7 color structure descriptor, run-length matrix statistics and spatial pyramid representation of these descriptors) and several textual models (i.e., probabilistic model, term frequency-inverse document frequency — TF-IDF; weighting and social TF-IDF weighting — an adaptation to the social space).

4. GROUND TRUTH

For each location, photos were manually annotated for relevance and diversity. Ground truth was generated by a small group of expert annotators with advanced knowledge of location characteristics. Software tools were specifically designed to facilitate the annotation process. The annotation process was not time restricted.

For *relevance annotation*, annotators were asked to label each photo (one at a time) as being relevant (value 1), non-relevant (0) or with "don't know" (-1). To help with their decisions, annotators were recommended to consult any additional information source during the evaluation (e.g., from the Internet). Final ground truth was determined after a majority voting scheme. The *devset* was annotated by 6 persons. The average inter-annotator agreement (Weighted Kappa) for the annotations of the *keywords* data was 0.68 ($\sigma = 0.07$) and for *keywordsGPS* data was 0.61 ($\sigma = 0.08$). The *testset* was annotated by 7 persons, each expert annotated a different part of the data set leading in the end to 3 annotations per image. The average inter-annotator agreement (Free-Marginal Multirater Fleiss' Kappa) for the annotation of the *keywords* data was 0.86 and for *keywords-GPS* data was 0.75.

Diversity annotation was carried out only for the photos that were judged as relevant in the previous step. For each location, annotators were provided with a thumbnail list of all relevant photos. After getting familiar with their content, they were asked to re-group the photos into similar visual appearance clusters (up to 20) and then tag these clusters with appropriate keywords. The *devset* was annotated by 3 persons and the *testset* by 4. In this case, each person annotated distinct parts of the data leading to only one annotation in the end.

To explore differences between expert and non-expert annotations, an additional crowd-sourcing annotated relevance and diversity ground truth was generated for a selection of 50 locations via CrowdFlower platform[4].

5. RUN DESCRIPTION

Participants were allowed to submit up to 5 runs. The first 3 are *required runs*: *run1* - automated approaches using visual information only; *run2* - automated approaches using textual information only; and *run3* - automated approaches using textual-visual information fused without other resources than provided by the organizers. The last 2 runs are *general runs*: *run4* - human-based or hybrid human-machine approaches and *run5* - everything allowed including using data from external sources (e.g., Internet). For generating *run1* to *run4* participants are allowed to use only information that can be extracted from the provided data (e.g., provided content descriptors, content descriptors of their own, etc). This includes also the Wikipedia webpage of the locations provided via their links. For *run5* everything is allowed, from the method point of view and information sources.

6. EVALUATION

Performance is assessed for both diversity and relevance. The main evaluation metrics is cluster recall at X (CR@X) [3] — a measure that assesses how many different clusters from the ground truth are represented among the top X results provided by the retrieval system. Precision at X (P@X) and the harmonic mean of CR@X and P@X (i.e., F1-measure@X) are used as secondary metrics. P@X measures the number of relevant photos among the top X results. F1-measure@X combines CR@X and P@X and gives and overall assessment of both diversity and relevance. Participants were provided with these metrics computed at different cutoff points, namely X$\in \{5, 10, 20, 30, 40, 50\}$. The official ranking was computed for X=10 (CR@10, P@10, F1-measure@10).

7. CONCLUSIONS

The Retrieving Diverse Social Images Task provides participants with a comparative and collaborative evaluation framework for social image retrieval techniques with explicit focus on result diversification, relevance and summarization. Details on the methods and results of each individual participant team can be found in the working note papers of the MediaEval 2013 workshop proceedings.

Acknowledgments

This task is supported by the following projects: EXCEL POSDRU, CUbRIK[5], PROMISE[6] and MUCKE[7]. Many thanks to the task supporters for their precious help: Anca-Livia Radu, Bogdan Boteanu, Ivan Eggel, Sajan Raj Ojha, Oana Pleş, Ionuţ Mironică, Ionuţ Duţă, Andrei Purica, Macovei Corina and Irina Nicolae.

8. REFERENCES

[1] S. Rudinac, A. Hanjalic, M.A. Larson, "Generating Visual Summaries of Geographic Areas Using Community-Contributed Images", IEEE Trans. on Multimedia, 15(4), pp. 921-932, 2013.

[2] R.H. van Leuken, L. Garcia, X. Olivares, R. van Zwol, "Visual Diversification of Image Search Results", ACM Int. Conf. on World Wide Web, pp. 341-350, 2009.

[3] M.L. Paramita, M. Sanderson, P. Clough, "Diversity in Photo Retrieval: Overview of the ImageCLEF Photo Task 2009", ImageCLEF 2009.

[3] all the provided photos are under Creative Commons licenses of type 1 to 7 that allow redistribution (see `http://www.flickr.com/services/api/flickr.photos.licenses.getInfo.html/` and `http://creativecommons.org/`).

[4] `http://crowdflower.com/`

[5] `http://www.cubrikproject.eu/`

[6] `http://www.promise-noe.eu/`

[7] `http://www.chistera.eu/projects/mucke/`