# Retrieving Diverse Social Images at MediaEval 2017: Challenges, Dataset and Evaluation

Maia Zaharieva[1], Bogdan Ionescu[2], Alexandru Lucian Gînscă[3],
Rodrygo L.T. Santos[4], Henning Müller[5]
[1]TU Wien, Austria
[2]University Politehnica of Bucharest, Romania
[3]CEA LIST, France
[4]Universidade Federal de Minas Gerais, Brazil
[5]University of Applied Sciences Western Switzerland, Switzerland
maia.zaharieva@tuwien.ac.at,bionescu@imag.pub.ro,alexandru.ginsca@cea.fr
rodrygo@dcc.ufmg.br,henning.mueller@hevs.ch

## ABSTRACT

This paper provides an overview of the Retrieving Diverse Social Images task that is organized as part of the MediaEval 2017 Benchmarking Initiative for Multimedia Evaluation. The task addresses the challenge of visual diversification of image retrieval results, where images, metadata, user tagging profiles, and content and text models are available for processing. We present the task challenges, the employed dataset and ground truth information, the required runs, and the considered evaluation metrics.

## 1 INTRODUCTION

An efficient image retrieval system should be able to present results that are both relevant to the provided query and that are covering different visual aspects of it. The diversification of image search results can considerably increase the probability of a system to address a broad range of user information needs. In general, diversification is an actively researched problem in various domains ranging from web search and query result diversification [9, 15, 19] to recommender systems [16, 17] and summarization [13, 14]. With the emerging availability of publicly available images, the importance for diversification of image data is steadily growing. This task is especially challenging when handling real-world queries, which are often complex, consisting of multiple concepts.

The 2017 Retrieving Diverse Social Images task is a follow-up of the 2016 edition [8] and fosters the development of new techniques for improving both the relevance and the visual diversification of image search results. The task is designed to support the evaluation and comparison of approaches emerging from a wide range of research fields, such as information retrieval (text, vision, multimedia communities), machine learning, relevance feedback, and natural language processing.

## 2 TASK DESCRIPTION

The task is built around the use case of a general ad-hoc image retrieval system, which provides the user with visually diversified representations of query results (see for instance Google Image Search[1]). Given a ranked list of up to 300 query-related images retrieved from Flickr[2] using text-based queries, participants are required to refine the results by providing a set of images that are *relevant* to the query and, at the same time, represent a *visually diversified* summary of it. The queries include complex and general-purpose, multi-concept queries (e.g. "dancing on the street", "trees reflected in water", "sailing boat"). The queries in the development set result from a broad user study and are constructed around the the data of the MediaEval 2016 Retrieving Diverse Social Images task [8]. The queries in the test set were collected using Google Trends[3] for image search (worldwide, last 5 years: 2012-2017).

The goal of the task is to refine the image set retrieved as a result to a given text-based query by providing a ranked list of up to 50 photos that are both *relevant* and *visually diversified* representations of the query, according to the following definitions:

**Relevance**: an image is considered to be relevant for the query if it is a common visual representation of the query topics (all at once). Bad quality photos (e.g., severely blurred, out of focus) are not considered relevant in this scenario;

**Diversity**: a set of images is considered to be diverse if it depicts different visual characteristics of the query topics and subtopics with a certain degree of complementarity, i.e. most of the perceived visual information is different from one image to another.

## 3 DATA DESCRIPTION

The data consists of a development set (*devset*) with 110 queries (32, 487 images) and a test set (*testset*) with 84 queries (24, 986 images). An additional dataset (*credibilityset*) provides credibility estimation for ca. 685 users and metadata for more than 3.5M images. We also provide semantic vectors for general English terms computed on top of the English Wikipedia[4] (*wikiset*), which could help participants to develop advanced text models.

Each query is accompanied by the following information: query text formulation (the actual query formulation used on Flickr to retrieve the data), a ranked list of up to 300 images in jpeg format retrieved from Flickr using Flickr's default "relevance" algorithm (all images are redistributable Creative Commons licensed[5]), an

*MediaEval'17, 13-15 September 2017, Dublin, Ireland*

[1] https://images.google.com/
[2] https://www.flickr.com.
[3] http://trends.google.com/
[4] https://en.wikipedia.org/
[5] http://creativecommons.org/

xml file containing Flickr metadata for the retrieved images, and ground truth for both relevance and diversity.

To facilitate participation from various communities, we also provide the following content-based descriptors:

- *general purpose, visual-based descriptors* extracted using the LIRE library[6] [10]: auto color correlogram (ACC) [12]; color and edge directivity descriptor (CEDD) [5], fuzzy color and texture histogram (FCTH) [6], Gabor texture, joint composite descriptor (JCD) [4], several MPEG7 features including color layout, edge histogram, and scalable color [11], pyramid of histograms of orientation gradients (PHOG) [2], and speeded up robust features (SURF) [1].

- *convolutional neural network (CNN)-based descriptors* based on the reference model provided with the Caffe framework[7]. The descriptors are extracted from the last fully connected layer (fc7).

- *text-based features* include term frequency and document frequency information and their ratio (TF-IDF). The text-based features are computed per image, per query, and per user basis.

- *user annotation credibility descriptors* provide an estimation of the quality of the users' tag-image content relationships. The following descriptors are provided: *visualScore* (measure of user image relevance), *faceProportion* (the percentage of images with faces), *tagSpecificity* (average specificity of a user's tags, where tag specificity is the percentage of users having annotated with that tag in a large Flickr corpus), *locationSimilarity* (average similarity between a user's geotagged photos and a probabilistic model of a surrounding cell), *photoCount* (total number of images a user shared), *uniqueTags* (proportion of unique tags), *uploadFrequency* (average time between two consecutive uploads), *bulkProportion* (the proportion of bulk taggings in a user's stream, i.e., of tag sets that appear identical for at least two distinct photos), *meanPhotoViews* (mean value of the number of times a user's image has been seen by other members of the community), *meanTitleWordCounts* (mean value of the number of words found in the titles associated with users' photos), *meanTagsPerPhoto* (mean value of the number of tags users put for their images), *meanTagRank* (mean rank of a user's tags in a list in which the tags are sorted in descending order according the number of appearances in a large subsample of Flickr images), and *meanImageTagClarity* (adaptation of the Image Tag Clarity from [18] using a TF-IDF language model as individual tag language model).

## 4 GROUND TRUTH

Both relevance and diversity annotations were carried out by 17 human annotators. The data were distributed among the annotators such that each query was labeled by three different annotators. For *relevance*, annotators were asked to label each image (one at a time) as being relevant to the underlying query (value 1), non-relevant (0), or with "don't know" ($-1$). The final relevance ground truth score was determined using a majority voting scheme. For *diversity*, only the images that were judged as relevant in the previous step were considered. For each query, annotators were provided with a thumbnail list of all relevant images. After getting familiar with their contents, they were asked to re-group the images into clusters with similar visual appearance (up to 25 clusters in total). In contrast to the single relevance score for each query, in terms of diversity we consider all three annotations as correct (ground truth) as they typically depict different possibilities to group the images representing different points of view.

## 5 RUN DESCRIPTION

Participants were allowed to submit up to five runs. The first three are required (dedicated) runs: *run1* – automated run using visual information only; *run2* – automated run using text information only; and *run3* – automated run using both visual and text information. For the generation of *run1* to *run3* only information that can be extracted from the provided data (e.g. provided descriptors, descriptors of their own, etc.) is allowed to be used. The last two runs, *run4* and *run5*, are general ones, i.e. any approach is allowed, e.g. human-based or hybrid human-machine approaches, including using data from external sources, such as Internet or pre-trained models obtained from external datasets related to this task.

## 6 EVALUATION

Performance is assessed for both diversity and relevance using *cluster recall* at $X$ ($CR@X$), *precision* at $X$ ($P@X$), and their harmonic mean $F1@X$. $CR@X$ provides the ratio of the number of clusters from the ground truth that are represented in the top $X$ results and, thus, it reflects the diversification quality of a given image result set. We compute $CR@X$ for each one of the available ground truth diversity annotations and select the one which maximizes $CR@X$ for each query. Since the clusters in the ground truth consider relevant images only, the relevance of the top $X$ results is implicitly measured by $CR@X$. Nevertheless, $P@X$ provides a more precise view on the relevance of a particular image set since it directly measures the relevance among the top $X$ images. We consider various cut off points, i.e. $X = \{5, 10, 20, 30, 40, 50\}$. Additionally, we consider two further evaluation metrics, which are well-established in the information retrieval community, the *intent-aware expected reciprocal rank* ($ERR\text{-}IA@X$) [3] and the *$\alpha$-normalized discounted cumulative gain* ($\alpha\text{-}nDCG@X$) [7] metrics. The official ranking metric is $F1@20$ which gives equal importance to diversity (via $CR@20$) and relevance (via $P@20$). This metric simulates the content of a single page of a typical Web image search engine and reflects user behavior, i.e., inspecting the first page of results with priority.

## 7 CONCLUSION

The 2017 Retrieving Diverse Social Images task provides participants with a comparative and collaborative evaluation benchmark for social image retrieval approaches focusing on *visual-based diversification*. The task explores the diversification in the context of a challenging, ad-hoc image retrieval system, which should be able to tackle complex and general-purpose multi-concept queries. This year, we explicitly accounted for the possibility of having multiple different views on a given retrieval result, which might all be subjectively correct. This allows for an investigation of the aspect of subjectivity in the perception of diversification in a next step. Details on the methods and results of the participating teams can be found in the working note papers of the MediaEval 2017 workshop proceedings.

[6] http://www.lire-project.net/

[7] http://caffe.berkeleyvision.org/

# REFERENCES

[1] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding* 110, 3 (2008), 346–359. https://doi.org/10.1016/j.cviu.2007.09.014

[2] Anna Bosch, Andrew Zisserman, and Xavier Munoz. 2007. Representing Shape with a Spatial Pyramid Kernel. In *ACM International Conference on Image and Video Retrieval (CIVR)*. ACM, New York, NY, USA, 401–408. https://doi.org/10.1145/1282280.1282340

[3] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *ACM Conference on Information and Knowledge Management (CIKM)*. ACM, New York, NY, USA, 621–630. https://doi.org/10.1145/1645953.1646033

[4] Savvas A Chatzichristofis, YS Boutalis, and Mathias Lux. 2009. Selection of the proper compact composite descriptor for improving content based image retrieval. In *Signal Processing, Pattern Recognition and Applications (SPPRA)*. ACTA Press, 134–140.

[5] Savvas A. Chatzichristofis and Yiannis S. Boutalis. 2008. CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval. In *International Conference on Computer Vision Systems (ICCV)*. Springer-Verlag, Berlin, Heidelberg, 312–322. https://doi.org/10.1007/978-3-540-79547-6_30

[6] S. A. Chatzichristofis and Y. S. Boutalis. 2008. FCTH: Fuzzy Color and Texture Histogram - A Low Level Feature for Accurate Image Retrieval. In *Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*. IEEE Computer Society, Washington, DC, USA, 191–196. https://doi.org/10.1109/WIAMIS.2008.24

[7] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and Diversity in Information Retrieval Evaluation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 659–666. https://doi.org/10.1145/1390334.1390446

[8] Bogdan Ionescu, Alexandru Lucian Ginsca, Maia Zaharieva, Bogdan Boteanu, Mihai Lupu, and Henning Müller. 2016. Retrieving Diverse Social Images at MediaEval 2016: Challenge, Dataset and Evaluation. In *MediaEval 2016 Multimedia Benchmark Workshop*, Vol. 1739. CEUR-WS.org.

[9] Bogdan Ionescu, Adrian Popescu, Anca-Livia Radu, and Henning Müller. 2016. Result diversification in social image retrieval: a benchmarking framework. *Multimedia Tools and Applications* 75, 2 (2016), 1301–1331. https://doi.org/10.1007/s11042-014-2369-4

[10] Mathias Lux. 2011. Content Based Image Retrieval with LIRe. In *ACM International Conference on Multimedia*. ACM, New York, NY, USA, 735–738. https://doi.org/10.1145/2072298.2072432

[11] B.S. Manjunath, J.-R. Ohm, V.V. Vasudevan, and A. Yamada. 2001. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology* 11, 6 (2001), 703–715. https://doi.org/10.1109/76.927424

[12] Mandar Mitra, Ramin Zabih, Jing Huang, Wei-Jing Zhu, and S. Ravi Kumar. 1997. Image Indexing Using Color Correlograms. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Washington, DC, USA, 762–768. https://doi.org/10.1109/CVPR.1997.609412

[13] Yanwei Pang, Qiang Hao, Yuan Yuan, Tanji Hu, Rui Cai, and Lei Zhang. 2011. Summarizing Tourist Destinations by Mining User-generated Travelogues and Photos. *Computer Vision and Image Understanding* 115, 3 (2011), 352–363. https://doi.org/10.1016/j.cviu.2010.10.010

[14] S. Rudinac, A. Hanjalic, and M. Larson. 2013. Generating Visual Summaries of Geographic Areas Using Community-Contributed Images. *IEEE Transactions on Multimedia* 15, 4 (2013), 921–932. https://doi.org/10.1109/TMM.2013.2237896

[15] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2015. Search result diversification. *Foundations and Trends in Information Retrieval* 9, 1 (2015), 1–90. https://doi.org/10.1561/1500000040

[16] Markus Schedl and David Hauger. 2015. Tailoring Music Recommendations to Users by Considering Diversity, Mainstreaminess, and Novelty. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 947–950. https://doi.org/10.1145/2766462.2767763

[17] Yue Shi, Xiaoxue Zhao, Jun Wang, Martha Larson, and Alan Hanjalic. 2012. Adaptive Diversification of Recommendation Results via Latent Factor Portfolio. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, 175–184. https://doi.org/10.1145/2348283.2348310

[18] Aixin Sun and Sourav S. Bhowmick. 2009. Image Tag Clarity: In Search of Visual-representative Tags for Social Images. In *SIGMM Workshop on Social Media*. ACM, New York, NY, USA, 19–26. https://doi.org/10.1145/1631144.1631150

[19] Kaiping Zheng, Hongzhi Wang, Zhixin Qi, Jianzhong Li, and Hong Gao. 2016. A survey of query result diversification. *Knowledge and Information Systems* 51, 1 (2016), 1–36. https://doi.org/10.1007/s10115-016-0990-4