Div400: A Social Image Retrieval Result Diversification Dataset

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

Anca-Livia Radu
DISI, University of Trento, Italy
radu@disi.unitn.it

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

Babak Loni
Delft University of Technology, The Netherlands
b.loni@tudelft.nl

ABSTRACT

In this paper we propose a new dataset, Div400, that was designed to support shared evaluation in different areas of social media photo retrieval, e.g., machine analysis (re-ranking, machine learning), human-based computation (crowdsourcing) or hybrid approaches (relevance feedback, machine-crowd integration). Div400 comes with associated relevance and diversity assessments performed by human annotators. 396 landmark locations are represented via 43,418 Flickr photos and metadata, Wikipedia pages and content descriptors for text and visual modalities. To facilitate distribution, only Creative Commons content was included in the dataset. The proposed dataset was validated during the 2013 Retrieving Diverse Social Images Task at the MediaEval Benchmarking Initiative for Multimedia Evaluation.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*collection, dissemination*.

Keywords

social photo retrieval, result diversification, multimedia content analysis, crowdsourcing, MediaEval benchmark, Flickr.

1. INTRODUCTION

Multimedia items make for an important share of the data distributed and searched for on the Internet. In particular, geographic queries represent a hefty chunk of users' queries. Current photo search technology is mainly relying on employing text, visual, or more recently on GPS information to provide users with accurate results for their queries. Retrieval capabilities are however still below the actual needs of the common user, mainly due to the limitations of the content descriptors, e.g., textual tags tend to be noisy or inaccurate (e.g., people may tag entire collections with a unique tag), automatic visual descriptors fail to provide high-level understanding of the scene while GPS coordinates capture the position of the photographer and not necessarily the position of the query. Until recently, research focused mainly on improving the *relevance* of the results. However, an efficient information retrieval system should be able to *summarize* search results so that it surfaces results that are both relevant and that are covering *different aspects* of a query (e.g., providing different views of a monument rather than duplicates of the same perspective).

In this paper we introduce a new dataset designed to support this emerging area of information retrieval that fosters new technology for improving both the relevance and diversification of search results with explicit focus on the actual social media context. This dataset is intended to support related areas of machine analysis (e.g., re-ranking, machine learning), human-based computation (e.g., crowdsourcing) as well as hybrid approaches (e.g., relevance feedback, machine-crowd integration).

The paper is organized as follows: Section 2 presents a brief overview of the literature and situates our contribution. Section 3 provides the description of the dataset while Section 4 discusses the annotation process. Section 5 provides insights about the validation of the dataset at the MediaEval 2013 benchmark and Section 6 concludes the paper.

MMSys'14, March 19 - 21 2014, Singapore

http://dx.doi.org/10.1145/2557642.2563670.

2. RELATED WORK

In the context of photo retrieval, relevance was more thoroughly studied than diversification and, even though a considerable amount of diversification literature exists, the topic remains a "hot" one, especially in social media [1, 2, 3, 4, 5]. One of the critical points of the diversification approaches are the evaluation tools. In general, experimental validation is carried out on closed datasets which limits the reproducibility of the results. Another weakness is the ground truth annotation which tends to be restrained, not enough attention being payed to its statistical significance. Contributions in this area are therefore highly valuable.

Closely related to our initiative is the ImageCLEF benchmarking and in particular the 2009 Photo Retrieval task [4] that proposes a dataset consisting of 498,920 news photographs (images and caption text) classified into sub-topics

(e.g., location type for locations, animal type for photos of animals) for addressing diversity. Other existing datasets are determined for the experimentation of specific methods. For instance [2] uses a collection of Flickr[1] images captured around 207 locations in Paris and as ground truth exploits the geographical coordinates accompanying the images. [3] addresses the diversification problem in the context of populating a knowledge base, YAGO[2], containing about 2 million typed entities (e.g., people, buildings, mountains, lakes, etc) from Wikipedia. [5] uses 75 randomly selected queries from Flickr logs for which only the top 50 results are retained; diversity annotation is provided by human assessors that grouped the data into similar appearance clusters.

The main contributions of the proposed dataset are: addressing the social dimension of the diversification problem that is reflected both in its nature (variable quality of photos and of metadata shared on social media) and in the methods devised to retrieve it — we use a tourist use case scenario where a person tries to find more information about a place she might visit and is interested in getting a more complete visual description of the place; providing ground truth annotation from both trusted assessors and crowd workers which will allow for exploring better the limits between the two as well as the annotation subjectivity; proposing a development framework (including evaluation tools) in the context of the actual retrieval technology as the dataset contains results obtained with Flickr's relevance system. These aspects are detailed in the sequel.

3. DATASET DESCRIPTION

Given the important role of geographic queries and their spatio-temporal and visual invariance, we created a dataset composed of tourist landmarks. The dataset consists of 396 landmark locations, natural or man-made, e.g., sites, museums, monuments, buildings, roads, bridges, houses, caves. They range from very famous ones, e.g., Big Ben in London, to lesser known to the grand public, e.g., Palazzo delle Albere in Italy. These locations were selected from the World Heritage Site of the United Nations Educational, Scientific and Cultural Organization (UNESCO)[3], the CUbRIK FP7 project[4] and the GEORAMA project[5] based on the number of redistributable photos available on Flickr.

The dataset consists of Creative Commons[6] Flickr and Wikipedia location data. For each location the following information is provided: *location name* (unique textual identifier in the dataset), *location id* (unique numeric identifier), *GPS coordinates* (latitude and longitude in degrees) retrieved from GeoHack[7] via the location Wikipedia web page, a link to its *Wikipedia web page*, a *representative photo* from Wikipedia, a *ranked set of photos* retrieved from Flickr (up to 150 photos), *metadata* from Flickr for all the retrieved photos and visual and text content descriptors[8].

[1] `http://www.flickr.com/`
[2] `http://datahub.io/dataset/yago/`
[3] `http://en.wikipedia.org/wiki/List_of_World_Heritage_Sites/`
[4] `http://www.cubrikproject.eu/`
[5] `http://georama-project.labs.exalead.com/`
[6] `http://creativecommons.org/`
[7] `http://tools.wmflabs.org/geohack/`
[8] to download the dataset see `http://traces.cs.umass.edu/index.php/mmsys/mmsys`.

3.1 Flickr data collection method

Apart from Wikipedia data, landmark information was collected from Flickr using Flickr API[9] (under Java) and the *flickr.photos.search* function. To compare different retrieval mechanisms, for some of the locations data was collected using only the location name as query while for the other part we used the name of the location and the GPS coordinates. For the text queries, data are retrieved by matching the provided keywords against the photo title, description or tags (parameter `&text=` location name). For the queries including the GPS coordinates, data is retrieved within a 1 Km radius around the provided coordinates (`&lat=` latitude in degrees, `&lon=` longitude in degrees, and `&radius=1`).

For each location, we retain, depending on their availability, at most the first 150 photo results (`&per_page=150`). All the retrieved photos are under Creative Commons licenses of type 1 to 7 that allow redistribution[9] (`&license=1,2,3,4,5,6,7`). For each photo, the retrieved metadata consist of the *photo's id* and *title*, photo *description* as provided by author, *tags*, geotagging information (*latitude* and *longitude* in degrees), the *date* the photo was taken, photo *owner's name*, the *number of times* the photo has been displayed, the *url link* of the photo location from Flickr[10], Creative Common *license type* (`&extras=description,tags,geo,date_taken,owner_name,views,url_b,license`), number of *posted comments* (via *flickr.photo.getInfo* function) and the photo's *rank* within the Flickr results (we generated a number from 1 to 150). Results were retrieved with Flickr's default "relevance" algorithm (`&sort=relevance`).

3.2 Visual and textual descriptors

The raw data retrieved from Wikipedia and Flickr is accompanied by automatically extracted visual descriptors and text models. Features are provided on an as-is basis with no guaranty of being correct.

3.2.1 Visual descriptors

For each photo, we provide the following descriptors:

- *global color naming histogram* (code *CN* — 11 values): maps colors to 11 universal color names [6];
- *global Histogram of Oriented Gradients* (code *HOG* — 81 values): represents the HoG feature computed on 3 by 3 image regions [7];
- *global color moments* on HSV (Hue-Saturation-Value) color space (code *CM* — 9 values): represent the first three central moments of an image color distribution: mean, standard deviation and skewness [8];
- *global Locally Binary Patterns* on gray scale (code *LBP* — 16 values) [9];
- *global Color Structure Descriptor* (code *CSD* — 64 values): represents the MPEG-7 Color Structure Descriptor computed on the HMMD (Hue-Min-Max-Difference) color space [10];
- *global statistics on gray level Run Length Matrix* (code *GLRLM* — 44 dimensions): provides 11 statistics computed on gray level run-length matrices for 4 directions: Short Run Emphasis, Long Run Emphasis, Gray-Level Non-uniformity, Run Length Non-uniformity, Run Percentage, Low Gray-Level Run Emphasis, High Gray-Level Run Emphasis, Short

[9] `http://www.flickr.com/services/api/flickr.photos.licenses.getInfo.html/`
[10] please note that by the time you use the dataset some of the photos may not be available anymore at the same url.

Table 1: Dataset image statistics.

	devset			testset		
	#locations	*#images*	*min-avg.-max #img./location*	*#locations*	*#images*	*min-avg.-max #img./location*
keywords	25	2,281	30 - 91.2 - 150	135	13,591	30 - 100.7 - 150
keywordsGPS	25	2,837	45 - 113.5 - 150	211	24,709	35 - 117.1 - 150
overall	50	5,118	30 - 102.4 - 150	346	38,300	30 - 110.7 - 150

Run Low Gray-Level Emphasis, Short Run High Gray-Level Emphasis, Long Run Low Gray-Level Emphasis, Long Run High Gray-Level Emphasis [11];
• *local spatial pyramid representations* (code 3x3) of each of the previous descriptors (image is divided into 3 by 3 non-overlapping blocks and descriptors are computed on each patch — the global descriptor is obtained by the concatenation of all values).

3.2.2 Text models

Text models were created using *tag* and *title* words of the Flickr metadata. Preprocessing consist of excluding English stop words and single character words. We provide the following models:
• *probabilistic model* (code *probabilistic*): estimates the probability of association between a word and a given location by dividing the probability of occurrence of the word in the metadata associated to the location by the overall occurrences of that word [12];
• *TF-IDF weighting* (code *tfidf*): term frequency-inverse document frequency is a numerical statistic which reflects how important a word is to a document in a collection or corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others [13];
• *social TF-IDF weighting* (code *social-tfidf*): is an adaptation of TF-IDF to the social space (documents with several identified contributors). It exploits the number of different users that tag with a given word instead of the term count at document level and the total number of users that contribute to a document's description. At the collection level, we exploit the total number of users that have used a document instead of the frequency of the word in the corpus. This measure aims at reducing the effect of bulk tagging (i.e., tagging a large number of photographs with the same words) and to put forward the social relevancy of a term through the use of the user counts [14]. All three models use the entire dataset to derive term background information, such as the total number of occurrences for the probabilistic model, the inverse document frequency for tf-idf or the total number of users for social-tfidf.

3.3 Dataset basic statistics

The landmarks in the dataset are unevenly spread over 39 countries around the world. The dataset is divided into a development set (code **devset**) containing 50 of the locations and whose objective is to serve for the design and training of potential approaches, and a testing dataset containing the remaining 346 locations (code **testset**) which is intended for validating the methods. Each of these datasets contain both data that was retrieved using only keywords (code **keywords**) and data retrieved with keywords and GPS coordinates (code **keywordsGPS** — see also Section 3.1). Some basic image statistics are presented in Table 1. Overall, the dataset contains some 43,418 images together with their metadata and descriptor information.

3.4 Data format

Each dataset is stored in an individual folder (*devset* and *testset*) containing on its turn individual folders for each category of data (*keywords* and *keywordsGPS*). Then, each dataset sub-folder contains the following information:
• **a topic xml file**: containing the list of the locations in the current dataset (e.g., *devsetkeywords_topics.xml* for devset keywords[11]). Each location is delimited by a <topic> </topic> statement and includes the location id, the name of the location, the GPS coordinates and the url to the Wikipedia webpage of the location;
• **an img folder**: containing all the retrieved Flickr images for all the locations in the dataset, stored in individual folders named after each location. Images are named after the Flickr photo ids. All images are stored in JPEG format and have a resolution of around 640 × 480 pixels;
• **an imgwiki folder**: containing Creative Commons location photos from Wikipidia (one photo per location[12]). Each photo is named after the location name and has the owner's name specified in brackets (e.g, "Basilica of Saint Peter Vatican (Wolfgang Stuck).jpg", author Wolfgang Stuck);
• **a xml folder**: containing all the Flickr metadata stored in individual xml files. Each file is named according to the location name and is structured as in the following example:
<photos **monument**="*Basilica of St Mary of Health Venice*">
<photo **date_taken**="*2003-10-09 15:19:30*" **description**="*The Basilica di Santa Maria della Salute, commonly known as ...*" **id**="*2323 210481*" **latitude**="*0*" **license**="*3*" **longitude**="*0*" **nbComments**="*50*" **rank**="*1*" **tags**="*europe italy venice ... sony f717*" **title**="*Basilica of St Mary of Health/Salvation, Venice*" **url_b**="`http://static.flickr.com/2410/2323210481_31da0f0311_b.jpg`" **username**="*Christopher Chan*" **views**="*6824*"/> ... </photos>

The *monument* value is the unique location name, then, each of the photos are delimited by a <photo /> statement. Each field is explained in Section 3.1;
• **a gt folder**: containing all the dataset ground truth files (more details are presented in Section 4);
• **a descvis folder**: containing all the visual descriptors. The *img* subfolder contains the descriptors for the Flickr images as individual csv (comma-separated values) files on a per location and descriptor type basis. Each file is named after the location name followed by the descriptor code, e.g., "Abbey of Saint Gall CM3x3.csv" refers to the global Color Moments (CM) computed on the spatial pyramid 3x3 for the location Abbey of Saint Gall (see Section 3.2.1). Within

[11] for testset, there are additional topic files (ending with "_") which are not containing the locations with no relevant images in the ground truth (namely, ids 81, 298, 305 and 367).
[12] please note that some locations have no images available.

each file, each photo descriptor is provided on an individual line (ending with carriage return). The first value is the unique photo id followed by the descriptor values separated by commas. The *imgwiki* subfolder contains the descriptors for the Wikipedia images as individual csv files on a per data set and descriptor type basis. Each file is named according to the dataset followed by the descriptor code, e.g., "devsetkeywordsGPS-CM.csv" refers to the global Color Moments (CM) for the devset keywordsGPS. As in the previous case, within each file, each Wikipedia photo descriptor is provided on an individual line (ending with carriage return). The first value is the Wikipedia photo file name followed by the descriptor values separated by commas;

• **a desctxt folder**: containing all the text models provided on a per dataset and model type basis. Each file is named according to the dataset followed by the descriptor code (see Section 3.2.2), e.g., "devsetkeywordsGPS-socialtfidf.txt" refers to the social TF-IDF weighting for the devset keywordsGPS. It contains a line for each location in the current dataset representing the model for that location (ending with carriage return). Columns are separated by tabulations. The first column of the line is the location name and the other columns contain each related word and its associated weight (associated words are sorted by decreasing scores).

4. DATASET ANNOTATION

The ground truth annotation of the dataset is strictly dependent on the use scenario intended for the dataset. As presented in Section 2, the proposed dataset was annotated in view of a tourist use case scenario where a person tries to find more information about a place she might visit. The dataset is annotated for relevance and diversity of the photos. The main annotation process was performed by experts (trusted annotators). To explore differences between expert and non-expert annotations, an additional crowd-sourcing annotation was generated for a selection of 50 locations from the *testset*. Dedicated visual software tools were employed to facilitate the process. The following definitions of relevance and diversity have been adopted:

• **relevance**: a photo is considered to be relevant for the location if it is a common photo representation of the location, e.g., different views at different times of the day/year and under different weather conditions, inside views, close-ups on architectural details, drawings, sketches, creative views, etc, which contain partially or entirely the target location. Bad quality photos (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 (see the examples above), with a certain degree of complementarity, i.e., most of the perceived visual information is different from one photo to another.

4.1 Annotations from trusted annotators

Trusted annotators have an advanced knowledge of the locations characteristics.

4.1.1 Task design

Relevance annotation task. For each location, the annotators were provided with one photo at a time. A reference photo of the location (e.g., the Wikipedia photo) was also displayed during the process. Annotators were asked to classify the photos as being relevant (score 1), non-relevant (0) or with "don't know" answer (-1). The definition of relevance was displayed to the annotators during the entire process. The annotation process was not time restricted. Annotators were recommended to consult any additional written or visual information source (e.g., from Internet) in case they were unsure about the annotation.

Diversity annotation task. Diversity is annotated only for the photos that were judged as relevant in the previous relevance step. For each location, annotators were provided with a thumbnail list of all the relevant photos. The first step required annotators to get familiar with the photos by analyzing them for about 5 minutes. Next, annotators were required to re-group the photos in clusters based on visual similarity. The number of clusters was limited to maximum 20. Full size versions of the photos were available by clicking on the photos. The definition of diversity was displayed to the annotators during the entire process. For each of the clusters, annotators provided also some keyword tags reflecting their judgments in choosing these particular clusters. The diversity annotation was also not time restricted.

4.1.2 Annotation statistics

The relevance ground truth for the *devset* was collected from 6 different expert annotators and the diversity one was collected from 3 experts that annotated distinct parts of the data set. For the *testset*, we employed 7 expert annotators that annotated different parts of the dataset leading in the end to 3 different annotations while the diversity ground truth was collected from 4 expert annotators that annotated distinct parts of the data set. Annotators were both females and males with ages ranging from 23 to 34. Final relevance ground truth was determined after a lenient majority voting scheme (equal numbers of 1 and 0 lead to a 1 decision, -1 are disregarded if not in majority).

The agreement among pairs of annotators was calculated using Kappa statistics, which measure the level of agreement discarding agreement given by chance. Kappa values range from 1 to -1, where values from 0 to 1 indicate agreement above chance, values equal to 0 indicate equal to chance, and values from 0 to -1 indicate agreement worse than chance. In general, Kappa values above 0.6 are considered adequate and above 0.8 are considered almost perfect [16].

Annotation statistics are summarized in Table 2 (for *devset* we report weighted Kappa and for *testset* Free-Marginal Multirater Fleiss' Kappa [16]). The relevance task statistics indicate a good agreement between annotators as well as the fact that retrieval using keywords and GPS data yields more accurate results than using solely the keywords. In total there were only 14 cases in which the majority voting is "don't know" (less than 0.04%, which is negligible). For the diversity annotation, overall, on average we obtain 11.6 clusters per location and 6.45 images per cluster for the *devset* and 13.2 and 5 for the *testset*, respectively.

4.2 Annotations from crowd workers

To explore differences between experts and non-experts annotations, the CrowdFlower[13] meta-crowdsourcing platform was used to annotate a subset of 50 locations from the *testset* (the actual list of locations is available in the *testsetcrowd_topics.xml* topic file, see Section 3.4 and 4.3).

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

Table 2: Expert annotation statistics.

devset		**testset**	
keywords	*keywordsGPS*	*keywords*	*keywordsGPS*
relevance (avg. Kappa / % relevant img.)			
0.68	0.61	0.86	0.75
68%	79%	55%	75%
diversity (avg. clusters per location / avg. img. per cluster)			
10.4	12.8	11.8	14.5
5.5	7.4	4.2	5.8

Table 3: Crowd annotation statistics.

testset (selection of 50 locations, 6169 photos)		
relevance (avg. Kappa and % relevant img.): 0.36 69%		
diversity (avg. clusters per location / avg. img. per cluster)		
GT1	*GT2*	*GT3*
3.5	4.3	6.3
43.1	30.4	24

4.2.1 Task design

Crowd-sourcing workers performed the *relevance* and *diversity* task annotations using the exact conditions described in Section 4.1.1, except for the fact that for the relevance annotation photos were annotated in sets of ten. Each set of pictures that was annotated for relevance was paid with 10 euro cents while for the diversity annotation workers were paid with 35 euro cents per location.

4.2.2 Quality control procedure

For the relevance task, the quality of the crowd-sourcing task was ensured using gold units. Gold unit is a quality control mechanism provided by CrowdFlower which consists in including unambiguous questions to select trusted annotations. Each annotator should at least answer four gold units with a minimum accuracy of 70% in order to be included in the set of trusted annotators. No trusted annotators are excluded from the final set of results. As recommended by CrowdFlower, 10% of the tasks were flagged as gold. For this purpose, a set of six additional locations and ten pictures related to each of them were collected. These locations were not included in the dataset. The set of collected pictures were unambiguously relevant or non-relevant.

Due to the subjective nature of the diversity task, gold units could not be used in a traditional way. Instead, an ad-hoc website was developed and made available in the crowd-sourcing task. Crowd-sourcing workers accessed the task via CrowdFlower. The CrowdFlower task contained the description of the task, an example of how to complete it, an open text field and a link to an external website which provided an interactive visual interface featuring drag and drop functionality to facilitate the annotations. After reading the description and looking at the example, crowd-sourcing workers were accessing the external interface. Here, they could cluster the set of relevant pictures on per individual location basis. Once all pictures were clustered, the interface provided a unique code that needed to be filled in the text field of the CrowdFlower task. In order to ensure quality, the open text field was flagged as gold unit and unique codes for all locations were provided in advance.

4.2.3 Annotation statistics

In total, 175 crowd-sourcing workers participated in the relevance task. On average, each worker performed 10.7 tasks (with a minimum of 3 and a maximum of 55). For each photo we retain three annotations. Final relevance ground truth was determined after the same lenient majority voting scheme as for trusted annotators (see Section 4.1.2). Annotation statistics are summarized in Table 3 (we report Free-Marginal Multirater Fleiss' Kappa [16]). In this case, one can observe that the agreement between annotators is significantly lower than for the trusted annotators (see Table 2) which proves the variable quality of crowd annotations. In total there were 62 cases in which the majority voting is "don't know" (around 1%). For the diversity task, there were in total 33 workers participating to the task. Workers performed an average of 11.8 tasks (with a minimum of 6 and a maximum of 24). We retain only three different annotations per location (selected empirically based on the coherence of the tags and number of clusters) for which, overall, on average we obtain 4.7 clusters per location and 32.5 images per cluster.

4.3 Annotation data format

Ground truth is provided on a per dataset, annotator type and location basis (see the folder structure in Section 3.4). Within the *gt* folder, relevance task ground truth is stored in the *rGT* subfolder and diversity task ground truth in the *dGT* subfolder. We provide individual txt files for each location. Files are named according to the location name followed by the ground truth code: *rGT* for relevance, *dGT* for diversity and *dclusterGT* for the cluster tags, e.g., "Abbey of Saint Gall dGT.txt" refers to the diversity ground truth for the location Abbey of Saint Gall.

For the *rGT* files, each file contains photo ground truth on individual lines (ending with carriage return). The first value is the unique photo id from Flickr followed by the ground truth value (1, 0 or -1) separated by comma. The *dGT* files are structured similarly to *rGT* but having after the comma the cluster id number to which the photo was assigned (a number from 1 to 20). The *dclusterGT* files, complements the *dGT* by providing the cluster tag information. Each line contains the cluster id followed by the cluster user tag separated by comma.

The crowd-sourcing annotation ground truth respects the formatting above but is stored in the *crowdsourcing* folder of *testset* together with its associated topic xml file (see Section 3.4). Each individual worker diversity ground truth is stored in a different subfolder (*dGT1* to *dGT3*).

5. MEDIAEVAL 2013 VALIDATION

The proposed dataset was validated during the 2013 Retrieving Diverse Social Images Task at the MediaEval Benchmarking Initiative for Multimedia Evaluation[17]. The task challenged participants to design either machine, human or hybrid approaches for refining Flickr results in view of providing a ranked list of up to 50 photos that are considered to be both relevant and diverse representations of the locations (for more details about the task see [15]).

In total, 24 teams from 18 countries registered to the task and 11 crossed the finish line. The proposed diversification approaches varied from graph representations, re-ranking, optimization approaches, data clustering to hybrid approaches that included human in the loop. Various combination of information sources have been explored. System performance was assessed in terms of cluster recall at X (CR@X — a measure that assesses how many different clusters from

Table 4: MediaEval 2013 results (top performance).

team & approach	ground truth	CR@10	P@10
SOTON-WAIS [17] & re-ranking (visual-text)	expert	43.98%	81.58%
	crowd	74.5%	77.14%
Flickr initial results	expert	36.49%	75.58%
	crowd	66.43%	68.16%

the ground truth are represented among the top X results), precision at X (P@X — measures the number of relevant photos among the top X results) and their harmonic mean, i.e., F1-measure@X (X∈{5,10,20,30,40,50}). Table 4 summarizes some of the best overall average results (for more details see the MediaEval workshop proceedings[17]). Highest performance for a cutoff at 10 images (used for the official ranking) was achieved using a re-ranking approach with a Greedy Min-Max similarity diversifier and using both visual and text information [17]. In terms of diversity, on average, it allows for an improvement close to 10% of Flickr initial ranking results (regardless the type of ground truth).

The following information will help reproducing the exact evaluation conditions of the task. Participant runs were processed in the form of trec topic files[14], each line containing the following information separated by whitespaces: *qid iter docno rank sim run_id*, where *qid* is the unique query id (see the topic files, Section 3.4), *iter* gets disregarded (e.g., 0), *docno* is the unique Flickr photo id, *rank* is the new photo rank in the refined list (an integer ranging from 0 — highest rank — up to 49), *sim* is a similarity score and *run_id* is the run label. A sample run file is provided in the root of the *testset* folder (*me13div_Example_run_visual_[dataset].txt*).

The official scoring tool, *div_eval.jar*, is available in the root of the *testset* folder. To run the script, use the following syntax (make sure you have Java installed on your machine): *java* **-jar** *div_eval.jar* **-r** <runfilepath> **-rgt** <rGT_path> **-dgt** <dGT _path> **-t** <topic_filepath> **-o** <output_dir> [optional: **-f** <filename>]; where <runfilepath> is the file path of the run file, <rGT_path> is the path to the relevance ground truth, <dGT_path> is the path to the diversity ground truth, <topic_filepath> is the file path to the topic xml file (official evaluation was carried out on all the locations with relevant pictures[11]).

6. CONCLUSIONS

We proposed a new dataset (Div400) that contains 396 landmark locations and 43,418 Creative Commons Flickr ranked photos together with their Wikipedia and Flickr metadata and some general purpose content descriptors (visual and text). Data was annotated for both the relevance of the results as well as for their diversity using trusted and crowd annotators. The dataset is intended for supporting research in areas related to information retrieval that focus on the diversification of search results and was successfully validated during the 2013 Retrieving Diverse Social Images Task at the MediaEval Benchmarking Initiative.

7. ACKNOWLEDGMENTS

This data set was supported by the following projects: EXCEL POSDRU, CUbRIK[4], PROMISE[15] and MUCKE[16]. We acknowledge the MediaEval Benchmarking Initiative for Multimedia Evaluation[17] and in particular Martha Larson and Gareth Jones for the very constructive discussions.

[14] http://trec.nist.gov/trec_eval/

[15] http://www.promise-noe.eu/

[16] http://www.chistera.eu/projects/mucke/

[17] http://www.multimediaeval.org/

8. REFERENCES

[1] A.-L. Radu, B. Ionescu, M. Menéndez, J. Stöttinger, F. Giunchiglia, A. De Angeli, "A Hybrid Machine-Crowd Approach to Photo Retrieval Result Diversification", Multimedia Modeling, Ireland, 2014.

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

[3] B. Taneva, M. Kacimi, G. Weikum, "Gathering and Ranking Photos of Named Entities with High Precision, High Recall, and Diversity", ACM Web Search and Data Mining, pp. 431-440, 2010.

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

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

[6] Van de Weijer, C. Schmid, J. Verbeek, D. Larlus, "Learning color names for real-world applications", IEEE Transactions on Image Processing, 18(7), pp. 1512-1523, 2009.

[7] O. Ludwig, D. Delgado, V. Goncalves, U. Nunes: "Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection", Intelligent Transportation Systems, 2009.

[8] M.Stricker, M. Orengo, "Similarity of color images", SPIE Storage and Retrieval for Image and Video Databases III, vol. 2420, pp. 381-392, 1995.

[9] T. Ojala, M. Pietikäinen, D. Harwood, "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions", IAPR Pattern Recognition, vol. 1, pp. 582-585, 1994.

[10] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, A. Yamada, "Color and texture descriptors", IEEE Transactions on Circuits and Systems for Video Technology, vol. 11(6), pp. 703-715, 2001.

[11] X. Tang, "Texture Information in Run-Length Matrices", IEEE Transactions on image Processing, vol.7(11), 1998.

[12] J.M. Ponte, W.B. Croft, "A Language Modeling Approach to Information Retrieval", Research and Development in Information Retrieval, pp. 275-281, 1998.

[13] G. Salton, M.J. McGill, "Introduction to modern information retrieval", McGraw-Hill, NY, USA, ISBN:0070544840, 1986.

[14] A. Popescu, G. Grefenstette, "Social Media Driven Image Retrieval", ACM ICMR, April 17-20, Trento, Italy, 2011.

[15] B. Ionescu, M. Menéndez, H. Müller, A. Popescu, "Retrieving Diverse Social Images at MediaEval 2013: Objectives, Dataset and Evaluation", MediaEval 2013 Workshop, Barcelona, Spain, October 18-19, 2013

[16] J.J. Randolph, "Free-Marginal Multirater Kappa (multirater κfree): an Alternative to Fleiss Fixed-Marginal Multirater Kappa", Joensuu Learning and Instruction Symposium, 2005.

[17] N. Jain, J. Hare, S. Samangooei, J. Preston, J. Davies, D. Dupplaw, P. Lewis, "Experiments in Diversifying Flickr Result Sets", MediaEval 2013 Workshop, Barcelona, Spain, October 18-19, 2013