Noname manuscript No.
(will be inserted by the editor)

Result Diversification in Social Image Retrieval: A Benchmarking Framework

Bogdan Ionescu · Adrian Popescu · Anca-Livia Radu · Henning Müller

Received: date / Accepted: date

Abstract This article addresses the diversification of image retrieval results in the context of image retrieval from social media. It proposes a benchmarking framework together with an annotated dataset and discusses the results achieved during the related task run in the MediaEval 2013 benchmark. 38 multimedia diversification systems, varying from graph-based representations, re-ranking, optimization approaches, data clustering to hybrid approaches that included a human in the loop, and their results are described and analyzed in this text. A comparison of the use of expert vs. crowdsourcing annotations shows that crowdsourcing results have a slightly lower inter-rater agreement but results are comparable at a much lower cost than expert annotators. Multimodal approaches have best results in terms of cluster recall. Manual approaches can lead to high precision but often lower diversity. With this detailed results analysis we give future insights into diversity in image retrieval and also for preparing new evaluation campaigns in related areas.

Keywords social photo retrieval · result diversification · image content description · re-ranking · crowdsourcing.

B. Ionescu
LAPI, University "Politehnica" of Bucharest, 061071, Romania,
E-mail: bionescu@alpha.imag.pub.ro

A. Popescu
CEA-LIST, Centre de Saclay - NanoInnov, France,
E-mail: adrian.popescu@cea.fr

A.-L. Radu
DISI, University of Trento, 38123 Povo, Italy,
LAPI, University "Politehnica" of Bucharest, 061071, Romania,
E-mail: ancalivia.radu@unitn.it

H. Müller
University of Applied Sciences Western Switzerland (HES-SO), Technopole 3, Sierre, Switzerland,
E-mail: henning.mueller@hevs.ch

1 Introduction

Multimedia such as videos and images make for an important share of the data distributed and searched for on the Internet. Current photo search technology is mainly relying on employing text annotations, 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., text tags tend to be inaccurate (e.g., people may tag entire collections with a unique tag) and annotation might have been done with a goal in mind that is different from the searchers goals. Automatically extracted visual descriptors often fail to provide high-level understanding of the scene [48] while GPS coordinates capture the position of the photographer and not necessarily the position of the query and they can again be assigned for a large set of images regardless of exact positions.

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 and give a global view 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 showing almost identical images. Relevance was more thoroughly studied in existing literature than diversification [1,4,5] and even though a considerable amount of diversification literature exists (mainly in the text-retrieval community), the topic remains important, especially in multimedia [7, 8,10–12].

Benchmarking activities provide a framework for evaluating systems on a shared dataset and using a set of common rules. The results obtained are thus comparable and a wider community can benefit from it. Campaigns such as TREC Video Retrieval Evaluation - TRECVID [2] (video processing and retrieval benchmarking), The CLEF Cross Language Image Retrieval - ImageCLEF [48] (image processing and retrieval benchmarking), The PASCAL Visual Object Classes - VOC [3] (pattern analysis benchmarking) or MediaEval Benchmarking Initiative for Multimedia Evaluation [19] (multimedia benchmarking) are well known successful examples of such initiatives. They contribute permanently to the actual scientific advances by challenging new techniques that can solve real-world processing requirements.

In this paper we introduce a new evaluation framework and dataset [32] (Div400) for benchmarking search result diversification techniques and discuss its contribution to the community by analyzing the results of the MediaEval 2013 Retrieving Diverse Social Images Task [31], which stood as validation. The proposed framework focuses on fostering new technology for improving both *relevance* and *diversification* of search results with explicit emphasis on the current social media context, where images are supplied and searched by members in social media platforms such as Flickr. These two characteristics of retrieval results are antinomic, i.e., the improvement of one usually results in a degradation of the other, which requires a deeper analysis.

We provide a comparative analysis of the state-of-the-art systems submitted to the task. Submitted systems addressed a broad category of approaches, varying from single-modal to multi-modal, from using graph representations, re-ranking, optimization approaches, clustering, heuristic to hybrid approaches that include humans in the loop. This analysis is helpful in that it evidences strong and weak points of current technology for diversifying social multimedia and can be used to guide further work in the area.

The remainder of the paper is organized as follows: Section 2 overviews the diversification literature and the evaluation frameworks. It situates the contribution of this paper compared to related work. Section 3 describes the proposed evaluation framework and dataset. Section 4 overviews the MediaEval 2013 participant systems. Before concluding, Section 5 provides a detailed analysis of the experimental results.

2 Previous work

The problem of retrieval results diversification was addressed initially for text based retrieval as a method of tackling queries with unclear information needs [18]. A typical retrieval scenario that focuses on improving the relevance of the results is based on the assumption that the relevant topics for a query belong to a single topic. However, this is not totally accurate as most of the queries involve many declinations such as for instance sub-topics, e.g., animals are of different species, cars are of different types and producers, objects have different shapes, points of interest can be photographed from different angles and so on. Therefore, one should consider equally the diversification in a retrieval scenario.

Improving the diversity of the results involves addressing the multiple possible intents, interpretations, or subtopics associated with a given query [18]. By widening the pool of possible results, one can increase the likelihood of the retrieval system to provide the user with information needed and thus to increase its efficiency. For instance, in user recommender systems, users will find satisfactory results much faster if the diversity of the results is higher [6].

A typical text-retrieval diversification approach involves two steps [17]. First, a ranking candidate set S with elements that are relevant to the user's query is retrieved. Second, a sub-set R of S is computed by retaining only the very relevant elements and at the same time a set that is as diverse as possible, i.e., in contrast to the other elements from the set R. The key of the entire process is to mitigate the two components (relevance and diversity — a bi-optimization process) which in general tends to be antinomic, i.e., the improvement of one of them usually results in a degradation of the other. Too much diversification may result in losing relevant items while increasing only the relevance will tend to provide many near duplicates.

Some of the most popular text diversification techniques explore Greedy optimization solutions that build the results in an incremental way (a review is presented in [17]). For instance, [13] describes an algorithm that relies on an

objective function computed on a probabilistic model. Relevance is achieved with standard ranking while diversity is performed through categorization according to a certain taxonomy. The aim is to find a set of documents that cover several taxonomies of the query. [15] uses a Greedy algorithm to compute a result set where diversification is achieved according to the document frequency in the data collection. Another example is the approach in [14] that uses absorbing Markov chain Random walks to re-rank documents. A document that was already ranked becomes an absorbing state, dragging down the importance of similar unranked states.

Transposed to multimedia items and more specifically in the context of social media, the diversification receives a new dimension by addressing multimodal (visual-text) and spatio-temporal information (video). Due to the heterogeneous nature of modalities, multimedia information is more complex and difficult to handle than text data. Assessing the similarity between multimodal entities has been one of the main research concerns in the community for many years. Common approaches are attempting to simplify the task by transposing the rich visual-text information into more simple (numeric) representations such as using content descriptors and fusion schemes. Diversification is then carried out in these multi-dimensional feature spaces with strategies that mainly involve machine learning (e.g., clustering).

Many approaches have been investigated. For instance, [7] addresses the visual diversification of image search results with the use of lightweight clustering techniques in combination with a dynamic weighting function of visual features to best capture the discriminative aspects of image results. Diversification is achieved by selecting a representative image from each obtained cluster. [47] jointly optimizes the diversity and the relevance of the images in the retrieval ranking using techniques inspired by Dynamic Programming algorithms. [10] aims to populate a database with high precision and diverse photos of different entities by revaluating relational facts about the entities. Authors use a model parameter that is estimated from a small set of training entities. Visual similarity is exploited using the classic Scale-Invariant Feature Transform (SIFT). [11] addresses the problem of image diversification in the context of automatic visual summarization of geographic areas and exploits user-contributed images and related explicit and implicit metadata collected from popular content-sharing websites. The approach is based on a Random walk scheme with restarts over a graph that models relations between images, visual features, associated text, as well as the information on the uploader and commentators. Also related to the social media context, it is worth bringing into discussion the accuracy and relevance of associated tags that help in the retrieval and diversification process. Current literature shows that some of the most efficient tag annotation and relevance assessment approaches rely on nearest-neighbors voting schemes, such as learning tag relevance via accumulating votes from visual neighbors proposed in [35] (an overview of these techniques is presented in [34]).

In the context of video data, the approach in [16] addresses representativeness and diversity in Internet video retrieval. It uses a video near-duplicate

graph that represents visual similarity relationship among videos on which near-duplicate clusters are identified and ranked based on cluster properties and inter-cluster links. Diversification is achieved by selecting a representative video from each ranked cluster (for a more comprehensive overview of the state-of-the-art see also Section 4).

Besides the scientific challenge, another critical point of the diversification approaches are the evaluation tools. In general, experimental validation is carried out on very particular and closed datasets, which limits the reproducibility of the results. Another weakness are the ground truth annotations that tend to be restrained and not enough attention is paid to their statistical significance and consequently to the statistical significance of the entire evaluation framework. There are however a few attempts to constitute a standardized evaluation framework in this area.

Closely related to our initiative is the ImageCLEF benchmarking and in particular the 2009 Photo Retrieval task [8] that proposes a dataset consisting of 498,920 news photographs (images and caption text) classified into subtopics (e.g., location type for locations, animal type for photos of animals) for addressing diversity. For assessing relevance and diversity authors propose the use of precision at cutoff at 20 images and instance recall at rank 20, which calculates the percentage of different clusters represented in the top 20 results. Evaluation is carried out on a total of 50 topics that were associated with a certain number of clusters.

Other existing datasets are determined for the experimentation of specific methods. For instance, [11] uses a collection of Flickr[1] images captured around 207 locations in Paris (100 images per location) to assess the diversity of visual summaries of geographic areas. Ground truth is, in this case, determined without the need of user input by exploiting the geographical coordinates accompanying the images, i.e., via an affinity propagation clustering of the latitude and longitude coordinates. Evaluation is performed using the outcome of a multinomial distribution which reaches its maximum when the relative number of geo-clusters' observations in the resulting image set corresponds to the clusters' prior probabilities (relative size of detected geo-clusters). [10] 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[3]. To assess performance, authors propose the use of standard Mean Average Precision (MAP) as well as of the Normalized Discounted Cumulative Gain (NDCG) — to measure the usefulness (gain) of images based on their (geometrically weighted) positions in the result list; and a preference-based measure, bpref, that does not depend on potential results (from the pool of all methods' results). Another example is the approach in [7] which uses 75 randomly selected queries from Flickr logs for which only the top 50 results are retained; diversity annotation is provided by

[1] `https://www.flickr.com/`
[2] `http://datahub.io/dataset/yago/`
[3] `http://http://en.wikipedia.org/`

human assessors that grouped the data into clusters with similar appearance. Performance is assessed using Folwkes-Mallows index-based metrics (the clustering equivalent of precision and recall — a high score of the Folwkes-Mallows index indicates that two clusterings are similar) and a criterion on variation of information (reduction of uncertainty from one clustering to the other).

As a general research trend in the field, methods operating on text data are now migrating and adapting to cope with the specificity of web multimedia information. Actually, different approaches perform differently on different types of data, e.g., text, image, video, and it is not always obvious to have clear positive and negative aspects of approaches. However, more and more focus is put on the actual social context with explicit focus on improving user satisfaction of the results.

In this paper we introduce a new evaluation framework and a dataset designed to support this emerging area of information retrieval that fosters new technology for improving both the relevance and diversification of search results. It proposes a dataset with 43,418 Flickr ranked photos of 396 geographic location landmarks that are annotated for both relevance and diversity. Annotations are carried out by experts as well as alternatively by crowd workers (for a part of the data). Diversification ground truth consists of regrouping images into similarity classes. An in-depth analysis of a selection of diversification approaches is reviewed as part of the experimental validation of this framework during the MediaEval 2013 Retrieving Diverse Social Images Task [31].

This work is a follow-up of our preliminary results presented in [32] and [33]. [32] is a dataset track paper that presents in detail the publicly available Div400 dataset with emphasis on physical data: resource structure and annotations. [33] is a short paper providing a brief overview of the MediaEval 2013 results, including the participant systems, precision vs. recall curves and comparison between expert and crowd annotations. Compared to our previous work, the main novelties of this paper are in the unified detailed description of both evaluation framework and dataset, in the in-depth analysis of the participant systems, in the extended analysis of the results from [33] and their implications and in addressing new experimental results that include results on a retrieval type basis (retrieval using keywords vs. GPS information), results on a location type basis, a detailed analysis of the stability statistics of the dataset and a user-based visual ranking experiment.

In the context of the current state-of-the-art the following main contributions of this work are identified:

- an evaluation framework is proposed that focuses on improving the current technology by using Flickr's relevance system as reference[4] (i.e., one of the state-of-the-art platforms) and addresses in particular the social dimension reflected in the nature of the data and methods devised to account for it;
- while smaller in size than the ImageCLEF collections [8,9], the proposed dataset contains images that are already associated to topics by Flickr.

[4] `http://www.flickr.com/services/api/`

This design choice ensures that there are many relevant images for all topics and pushes diversification into priority;
- unlike ImageCLEF, which worked with generic ad-hoc retrieval scenarios, a focused real-world usage scenario is set up, i.e. tourism, to disambiguate the diversification need;
- a comparison of expert and crowd-sourced ground truth production is performed to assess the potential differences between lab and real life evaluations;
- a comparative analysis of a broad variety of diversification approaches is proposed, varying from automatic to hybrid human-machine, that evidences strong and weak points of current diversification of social media technology and can be used to guide further work in the area.

3 Experiment and data description

To benchmark retrieval diversification techniques, the following task was designed and validated within the 2013 MediaEval benchmark [19]. The task builds on current state-of-the-art retrieval technology, e.g., using the Flickr media platform[4], with the objective of fostering approaches that will push forward the advances in the field.

3.1 Dataset

Given the important proportion of geographic queries and their spatio-temporal invariance, experimentation with the retrieval of photos with landmark locations was considered as this is a typical scenario that many users might be in. The proposed dataset consists of 396 landmark locations, natural or man-made, e.g., bridges, arches, cathedrals, castles, stadiums, gardens, monuments, that range from very famous ones, e.g., Big Ben in London, to monuments less known to the public, e.g., Palazzo delle Albere in Italy. Some examples are illustrated in Figure 1.

The locations are unevenly distributed around the world based on the availability of photos (see Figure 2): Arab Emirates (1 location), Argentina (12), Australia (2), Austria (3), Belgium (5), Brazil (4), Bulgaria (2), Cambodia (1), Canada (1), Chile (5), China (15), Colombia (4), Denmark (3), Egypt (2), France (48), Germany (20), Greece (4), Holland (1), India (19), Indonesia (1), Ireland (1), Italy (81), Japan (1), Jerusalem (1), Mexico (9), New Zealand (2), Pakistan (1), Paraguay (1), Peru (4), Portugal (5), Romania (4), Russia (2), Scotland (1), Spain (38), Switzerland (1), Turkey (3), United Kingdom (27), United States (60) and Venezuela (1).

For each location up to 150 photos (with Creative Commons redistributable licenses) and associated metadata are retrieved from Flickr and ranked with Flickr's default "relevance" algorithm. To compare different retrieval mechanisms, data were collected with both text (i.e., location name — *keywords*) and

Fig. 1: Example pictures from the dataset (photo credits from Flickr, from left to right and top to bottom: Andwar, Ipoh kia, Marvin (PA), photoAtlas, Julie Duquesne, Jack Zalium and kniemla).

Fig. 2: Location distribution (image form Google Maps ©2013 MapLink).

GPS queries (*keywordsGPS*). Location metadata consists of Wikipedia links to location webpages and GPS information. On the other hand, photo metadata include social data: *photo 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, Creative Common *license type*, number of *posted comments* and the photo's *rank* within the Flickr results (we generated a number from 1 to 150).

Apart from these data, to support contributions from different communities, some general purpose visual and text content descriptors are provided for the photos. Visual descriptors consist of global *color histograms*, global *Histogram of Oriented Gradients* (HoG), global *color moments* computed on HSV

(Hue-Saturation-Value) color space, global *Locally Binary Patterns* (LBP) computed on gray scale, global *MPEG-7 Color Structure Descriptor*, global *statistics on gray level Run Length Matrix*; together with their local spatial pyramid representations, i.e., images are 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.

Text descriptors include a probabilistic model that 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; TF-IDF weighting — term frequency-inverse document frequency that 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; and social TF-IDF weighting — 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, the total number of users that have used a document is exploited 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 relevance of a term through the use of the user counts [36]. 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-TF-IDF.

The dataset includes a total of 43,418 photos and is divided into a *devset* of 50 locations (5,118 photos, in average 102.4 per location) intended for designing and tuning the methods and a *testset* of 346 locations (38,300 photos, in average 110.7 per location) for the evaluation. The dataset is publicly available[5] — a description can be found in [31,32].

3.2 Ground truth annotation

The ground truth annotation of the dataset is strictly dependent on the use scenario intended for the dataset. As previously mentioned, 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 both relevance and diversity of the photos. The following definitions were 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

[5] data can be downloaded from `http://traces.cs.umass.edu/index.php/mmsys/mmsys/`

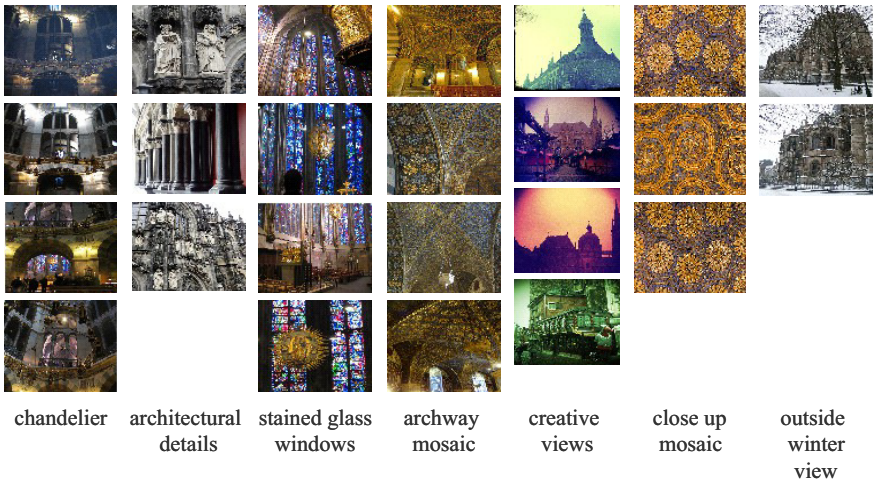

Fig. 3: Diversity annotation example for location "Aachen Cathedral" in Germany (excerpt from the total number of 15 clusters). The cluster tags are depicted at the bottom.

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 — photos are tagged as relevant, non-relevant or with "don't know" answer;

• **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 — relevant photos are clustered into visually similar groups and a tag that best describes the choice is provided to each cluster. An example is illustrated in Figure 3.

Definitions were determined and validated in the community based on the feedback gathered from 36 respondents during the 2013 MediaEval survey. MediaEval[6] is a "bottom-up benchmark" which means that the tasks that it offers are highly autonomous and are released based on the feedback gathered from the target community.

Annotations were carried out mainly by knowledgeable assessors (experts) with advanced knowledge of location characteristics (which was gained either via physical visits or via an in-depth exploration of the Internet and of the Flickr metadata). To avoid any bias, annotations were carried out individually on different locations without having the annotators discussing with each

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

other. To explore differences between expert and non-expert annotations, a subset of 50 locations from the *testset* was annotated using crowd-workers. In all cases, visual tools were employed to facilitate and ease the process (e.g., regrouping of photos is done using a graphical user interface that allows drag-and-drop facilities).

To assess the relevance of the ground truth, we present a detailed discussion on the annotation statistics. A summary of the overall statistics is presented in Table 1 that depicts the number of distinct annotations (the number of annotators is presented in the brackets), Kappa inter-annotator agreement and cluster statistics.

3.2.1 Devset annotation statistics

The *devset* relevance ground truth was collected from 6 expert annotators (with ages ranging from 23 to 34) and final ground truth was determined after a lenient majority voting scheme. The agreement among pairs of annotators was calculated using Kappa statistics, which measure the level of agreement among annotators discarding agreement given by chance. For this study, Weighted Kappa [37] was used. This variant of Cohen's Kappa [38] measures the level of agreement among annotators considering annotations having different weights. In particular, disagreement involving distant values (i.e., "relevant"/"non-relevant") are weighted more heavily than disagreements involving more similar values (i.e., "relevant"/"don't know"). Kappa values range from 1 to -1, with 0 indicating no correlation, -1 perfect negative correlation and 1 perfect correlation. As a general guideline [37], Kappa values above 0.6 are considered adequate and above 0.8 are considered almost perfect.

In our set of annotations, the average Kappa value for the expert annotations of images retrieved using only keywords was 0.68 (standard deviation 0.07), with minimum/maximum values equal to 0.56/0.8. The average Kappa value for the expert annotations of images retrieved using keywords and GPS coordinates was 0.61 (standard deviation 0.08), with minimum/maximum values equal to 0.49/0.75. After majority voting, 68% of the images retrieved using only keywords and 79% of the images retrieved using keywords and GPS coordinates were relevant. In total there were only 3 cases in which the majority voting is "don't know" (less than 0.06%).

The diversity ground truth was collected from 3 expert annotators that annotated distinct parts of the data set. It leads to an average number of 11.6 clusters per location and 6.4 images per cluster. Overall results are presented in Table 1.

3.2.2 Testset annotation statistics

The *testset* relevance ground truth was collected from 7 expert annotators (with ages ranging from 23 to 34). Each expert annotated a different part of the data set leading in the end to 3 different annotations for the entire data

Table 1: Expert and crowd annotation statistics.

devset (expert)	*testset* (expert)	*testset* (crowd)
relevance (annotations - avg.Kappa - % relevant img.)		
6(6) - 0.64 - 73	3(7) - 0.8 - 65	3(175) - 0.36 - 69
diversity (annotations - avg.clusters/location - avg.img./cluster)		
1(3) - 11.6 - 6.4	1(4) - 13.1 - 5	3(33) - 4.7 - 32.5

set. As for the devset, final ground truth was determined after a lenient majority voting scheme. For this study, Free-Marginal Multirater Fleiss' Kappa [39] was used. This Kappa statistic is appropriate for annotation tasks with more than two raters who annotate parts of the data set. The free-marginal variation is recommended, instead of the fixed-marginal version, when annotations' distribution among cases is not restricted [39].

In our set of annotations, the Kappa value for the expert annotations of images retrieved using only keywords was 0.86 and the Kappa value of the annotations of images retrieved using keywords and GPS coordinates was 0.75. After majority voting, 55% of the images retrieved using only keywords and 75% of the images retrieved using keywords and GPS coordinates were relevant. In total there were only 14 cases in which the majority voting is "don't know" (less than 0.04%).

The diversity ground truth was collected from 4 expert annotators that annotated distinct parts of the data set. It leads to an average number of 13.1 clusters per location and 5 images per cluster — see overall results in Table 1.

3.2.3 Crowdset annotation statistics

To explore differences between experts and non-experts annotations, the CrowdFlower[7] meta-crowdsourcing platform was used to annotate a subset of 50 locations from the *testset*. Crowdsourcing workers performed the relevance and diversity task annotations using the exact conditions as for the expert annotations, except for the fact that for the relevance annotation, photos were annotated in packs of ten (instead of having the entire set). 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.

For the relevance task, the quality of the crowdsourcing task was ensured using gold units. Gold unit is a quality control mechanism provided by CrowdFlower which consists of 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. Non-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

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

them were collected. These locations were not included in the dataset. The set of collected pictures is unambiguously relevant or non-relevant.

In total, 175 crowdsourcing 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. In this case, the agreement between annotators is significantly lower than for the trusted annotators (see also Table 1), i.e., 0.36 (Free-Marginal Multirater Fleiss' Kappa [39]), which may reflect the variability of the background of the crowd annotators. After majority voting, 69% of the images were relevant. 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 in 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.

3.3 Experiment

Given the dataset described above, the benchmarking requires developing an approach that allows the refinement of the initial Flickr retrieval results to retain only a ranked list of up to 50 photos that are equally relevant and diverse representations of the query (the number 50 was selected in view of addressing a limited number of results that would fit into a typical page of search results). This will require filtering the results, as initial retrieval results are inaccurate, e.g., depicting people in focus, other views or places, meaningless objects present at the location; as well as diversify them to reduce their redundancy, e.g., remove photo duplicates or similar views that provide the same visual information.

4 System descriptions

In this section we overview the 11 systems tested on the proposed evaluation framework during the MediaEval 2013 Retrieving Diverse Social Images Task [31] with the objective of validating the framework. Diversification approaches varied from graph-based representations, re-ranking, optimization approaches, data clustering to hybrid approaches that included a human in the loop. Various combination of information sources have been explored: visual, textual, multimodal or human-machine:

– **SOTON-WAIS** (*re-ranking, Greedy optimization — multimodal*) [20]: uses a three step approach that involves pre-filtering of the initial results,

re-ranking and a Greedy Min-Max diversifier. Pre-filtering is used to improve precision and attempts to remove images unlikely to be relevant, i.e., items which contained frontal or side-views of faces in focus, blurred, out-of-focus, with high amount of text, geotagged more than 8 km away from the actual location, without any views on Flickr or containing a description over 2,000 characters long. For the runs that included text and metadata, a proximity-based re-ranking is employed via performing a phrase or proximity search in which the results are scored higher if the query terms occur in close proximity in the metadata (use of Lucene[8]). Finally, diversification is carried out with a Greedy Min-Max diversifier that takes as input a similarity matrix between images and a pivot image. The pivot image is taken as the first image in the results, the second image is the one that has minimum similarity to the pivot and the remaining images are then selected such that they have the maximum dissimilarity to all of the previously chosen images;

– **SocSens** (*Greedy optimization, clustering — multimodal, human*) [25]: the visual approach involves the Greedy optimization of a utility function that weights both relevance and diversity scores. In particular, ground truth data (devset) was used to train a classifier whose prediction for an image is used as relevance score. The diversity score is defined as the dissimilarity of the current image to the most similar image from the set (Euclidean distance between the VLAD+SURF vectors [42] of the images is used as metric).

  The first text-based approach involves Hierarchical Clustering with image ranking using random forests. Diversification is achieved by stepping through the clusters iteratively and selecting the most relevant images until the requested number of images is achieved. A second text-based approach uses camera Exif information (i.e., date and time the photo was taken, aperture size and exposure time to determine whether the picture is indoor or outdoor, geo-location data that is used to determine the angle and distance to the photographed landmark) and weather data (i.e., weather conditions of the day the picture was taken) with k-means clustering to diversify images based on distance from the landmark, angle of the shot, weather conditions and time of the day.

  To leverage both visual and text information, a multimodal approach involves the late fusion of the outputs of the previous schemes. This is implemented by taking the union of the images returned for each location by the two previous approaches and ordering them in ascending average rank.

  Finally, a hybrid human-machine approach combines human and computer responses. A number of human assessors were provided with computer-generated short-lists of images (limited to 15 images) and asked to select 5 poor-quality or nearly duplicate images. The refined results consist of

[8] `http://lucene.apache.org/`

the 10 remaining images followed by the rejected ones (the numbers are selected thus to adapt to the official ranking that was set to a cutoff of 10 images);

- **MUCKE** (*re-ranking, cluster ranking — multimodal*) [27]: uses a two step approach in which images are first re-ranked to remove noisy items and then clustered to diversify results. A k-Nearest Neighbour inspired algorithm is proposed in which images of the target location constitute the positive set and a sample of images from the other locations in the dataset is used as a negative set. GIST is used to describe the content of the each positive image and its rank is given by counting the different users which upload positive images among the first 5 visual neighbours. Ties are broken by looking at the average distance between the image and its top 5 positive neighbours. Only the first 70% of the re-ranked images are retained for the diversification step. kMeans++ is used to cluster the remaining images and a number of 15 clusters is retained. To maximise their social relevance, these clusters are ranked by the number of different items that contribute to them, by the number of different dates when photos were taken and finally by the total number of included images. The final list of results is created by iteratively selecting images from the top 10 clusters;

- **CEA** (*re-ranking, social cues, informativeness — multimodal*) [29]: focuses on the use of social cues (user and temporal information) in the retrieval process and on their combination with visual cues. The diversification relies on the use of an informativeness measure which accounts for the novelty brought by each candidate with respect to candidates which were already selected. The simplest runs exploit the initial Flickr ranking and diversify images iteratively by selecting, in each round, images which are new using a social criterion that can be either the user ID or, in a more relaxed version, the user ID and the date of the image. The same algorithm used by the MUCKE [27] team is used to obtain an initial re-ranking of images, the only difference being that visual content is described using HoG instead of GIST. In the diversification step, new images are selected by maximizing their visual distance to the images which were already selected;

- **UPMC** (*re-ranking, clustering — multimodal*) [21]: uses re-ranking to improve relevance. To compare images, several similarity metrics are used, e.g., Euclidean distance for visual descriptors, Dirichlet Prior Smoothing and cosine for textual models, classical great-circle distance Haversine formula for the distance between two GPS coordinates. In addition, to better exploit geographical granularity between images, the Xilopix thesaurus[9] is used to convert image information into concepts (i.e., by matching the query GPS coordinates or keywords to thesaurus concepts). Similarity between concepts is evaluated using Wu-Palmer's similarity [21]. After re-

[9] `http://media-manager.xilopix.com/`

ranking, an Agglomerative Hierarchical Clustering is used to regroup images into similar appearance clusters. Diversification is achieved by cluster image sorting according to a priority criterion (e.g., decreasing the number of images in the cluster) and a final re-ranking that alternates images from different clusters;

– **MMLab** (*clustering, Greedy optimization — multimodal*) [26]: uses for the visual approach a variant of LAPI's approach [23]. Different from [23], the proposed approach uses a preliminary hierarchical clustering of the images while image similarity is assessed with a Gaussian kernel function.

  The text-based approach uses textual relevance and semantic similarity to diversify the results. It relies on a Greedy optimization of an estimate of the Average Diverse Precision metrics (variant of classical Average Precision that takes into account also the diversity). Relevance estimation of images is modeled as a linear combination of several information sources, such as the number of views, number of comments and textual models for the tags. Diversity estimation for an image is determined based on the minimal difference against the other images.

  The multimodal approach uses the text-based approach to estimate the relevance while similarity between images is determined using visual information. To account for both relevance and diversity, Hierarchical Clustering is employed as a final step (candidate images are selected as cluster representative images);

– **BMEMTM** (*heuristic, clustering, re-ranking — multimodal, human*) [24]: uses a visual approach that clusters the images using Hierarchical Clustering. Diversification is achieved by re-ranking initial results thus to output images from different clusters. Prior to the clustering, a face detector is used to downrank images containing faces.

  The text-based diversification consists mainly of a re-ranking scheme that uses the weights from the provided text models. Each image is associated with a score from each text model (i.e., the sum of the maximum values from all the keywords related to the image and the logarithm of the image average value). Besides the provided models (see Section 3.1), three improved variants are considered — tags without spaces (e.g., "basilicadisantamariadellasalute") are split into keywords and models re-computed. Final diversification is achieved by determining the ranks based on an average weight score (a higher score means a higher rank).

  To account for multi-modality, the previous two approaches are used in cascade: text-based followed by visual.

  A human-based approach included the user in the loop. A specific tool was designed to allow users to cluster the images and tag their relevance. Once input was collected, final ranking is determined by progressively selecting and removing from each non-empty user generated cluster the most rele-

vant images. Selected images are ordered according to their initial Flickr rank;

– **TIA-INAOE** (*functional optimization — multimodal*) [22]: transposes the diversification problem into the optimization of an objective function that combines relevance and diversity estimates. The target is to determine a ranked list that provides the best tradeoff between Spearman's correlation coefficient [40] of the difference between the initial ranking and the target one and a diversity coefficient that assesses the visual similarity between images (images are represented with content descriptors). Optimization is achieved with a multi-objective evolutionary algorithm, NSGA-II, that is able to simultaneously maximize the two previous measures, i.e., the image diversity in consecutive positions while minimizing divergence from the original ranking;

– **LAPI** (*re-ranking, clustering — multimodal*) [23][12]: uses a re-ranking with a clustering approach to select from results a small set of images that are both relevant and diverse representations of the query. First, a similarity matrix is obtained by computing for each image the average distance to the remaining images (images are represented with content descriptors). Then, a Synthetic Representative Image Feature (SRI) is determined by averaging the array. To account for relevance, a relevance rank is obtained by sorting images according to the new similarity values obtained after subtracting the SRI value from the similarity array. Furthermore, re-ranked images are clustered using a k-means approach. For each cluster a new SRI is estimated and a new re-ranking is performed. Each cluster is then represented by selecting only the top ranked images and again a similarity array is built. To account for diversity, a diversity rank is obtained by sorting in descending order the new similarity values corresponding to the previously selected images. Final ranking of the images is achieved by averaging relevance and diversity ranks and sorting them in ascending order;

– **UEC** (*web inspired ranking — multimodal*) [28]: uses an adaptation of VisualRank [43] (rank values are estimated as the steady state distribution of a random-walk Markov model) for improving precision followed by ranking with Sink Points for diversification [44]. First, a similarity matrix between images is determined using content descriptors. VisualRank is then applied to determine the most representative photo (i.e., the one ranked first). The remaining images are re-ranked by ranking with Sink Points. The process is repeated by extracting at each step the top ranked images until the target number of photos is achieved;

– **ARTEMIS** (*graph representation — visual*) [30]: exploits solely the representative power of visual information with a graph-based representation approach. Images are represented with Bag-of-Visual-Words of Hessian-Affine co-variant regions and RootSIFT descriptors. A landmark matching graph

is constructed where images are nodes and edges connect similar images. Multiple instances of a given landmark are identified as connected components in the landmark graph, and from each such component the dominant images are chosen as being representative. Diversification is achieved by selecting from each cluster the images with the highest similarity scores cumulated over its matches.

5 Experimental results

This section presents the results achieved during the 2013 MediaEval Retrieving Diverse Social Images Task [31] which received 38 runs from 11 participant teams. During the competition, participants designed and trained their methods on the *devset* dataset (50 locations and 5,118 photos) while the actual benchmarking was conducted on the *testset* (346 locations and 38,300 photos; see also Section 3). Participants were allowed to submit the following types of runs: automated techniques that use only visual information (*run1*), automated techniques that use only text information (*run2*), automated techniques that use multimodal information fused without other resources than provided (*run3*), human-based or hybrid human-machine approaches (*run4*) and a general run where everything was allowed including using data from external sources like the Internet (*run5*).

Performance is assessed for both diversity and relevance. The main evaluation metric was chosen to be Cluster Recall at X ($CR@X$) [8], defined as:

$$CR@X = \frac{N}{N_{gt}} \tag{1}$$

where N is the number of image clusters represented in the first X ranked images and N_{gt} is the total number of image clusters from the ground truth (N_{gt} is limited to a maximum of 20 clusters from the annotation process). Defined this way, $CR@X$ assesses how many clusters from the ground truth are represented among the top X results provided by the retrieval system. Since clusters are made up of relevant photos only, relevance of the top X results is implicitly measured by $CR@X$, along with diversity.

However, to get a clearer view of relevance, Precision at X ($P@X$) is also used as a secondary metric and defined as:

$$P@X = \frac{N_r}{X} \tag{2}$$

where N_r is the number of relevant pictures from the first X ranked results. Therefore, $P@X$ measures the number of relevant photos among the top X results. To account for an overall assessment of both diversity and precision, $F1@X$ was also reported which is the harmonic mean of $CR@X$ and $P@X$:

$$F1@X = 2 \cdot \frac{CR@X \cdot P@X}{CR@X + P@X} \tag{3}$$

Table 2: Keywords vs. keywords and GPS retrieval - Flickr initial results.

dataset	metrics	@5	@10	@20	@30	@40	@50
keywords	P	0.7682	0.7045	0.6602	0.6364	0.6169	0.5856
keywordsGPS		0.801	0.7881	0.7721	0.7716	0.7652	0.7518
keywords	CR	0.2618	0.3985	0.5745	0.6855	0.767	0.8113
keywordsGPS		0.215	0.3437	0.5095	0.6371	0.7249	0.791
keywords	F1	0.3685	0.4826	0.5824	0.6254	0.6456	0.6397
keywordsGPS		0.3282	0.461	0.593	0.6778	0.7241	0.7485

Evaluation was conducted for different cutoff points, X$\in\{5,10,20,30,40,50\}$. In particular, submitted systems were optimized with respect to CR@10 (i.e., for 10 images returned) which was the official metric. CR@10 was chosen because it ensures a good approximation of the number of photos displayed on different types of screens and also in order to fit the characteristics of the dataset (at most 150 images and 20 clusters per location). It is worth mentioning that given the definition in equation 1, CR@10 is inherently limited to a highest possible value of 0.77, as on average the dataset has 13 clusters per location (see Table 1).

Results are presented in the following sections. We report the average values over all the locations in the dataset.

5.1 Keywords vs. keywords and GPS retrieval

The first experiment consists of assessing the influence of the query formulation method on the results. For experimentation, we consider Flickr's default "relevance" algorithm that was used to collect the data (initial results) and which constitutes our baseline for comparing the systems' results. As presented in Section 3, *testset* data was collected with two approaches: using only the location name as query (*keywords* — 135 locations, 13,591 images) and using both the name of the location and its GPS coordinates (*keywordsGPS* — 211 locations, 24,709 images). For the text queries, data are retrieved by Flickr by matching the provided keywords against the photo title, description or tags. For the queries including the GPS coordinates, data is retrieved within a 1 Km radius around the provided coordinates. Table 2 summarizes the results achieved for the two approaches on the *testset*. We report location-based averages over each subset.

As expected, retrieval including GPS information yields more accurate results than using solely keywords, e.g., for the initial Flickr results, P@10 with keywords is 0.7045 compared to 0.7881 using GPS data. The differences between the two tend to increase with the number of images as the probability of including non-relevant images is higher, e.g., P@50 with keywords is 0.5856 compared to 0.7518 using the GPS information. On the other hand, diversity tends to be slightly higher for keywords, e.g., CR@10 is 0.3985 compared to 0.3437 using GPS. By increasing the number of images this difference tends

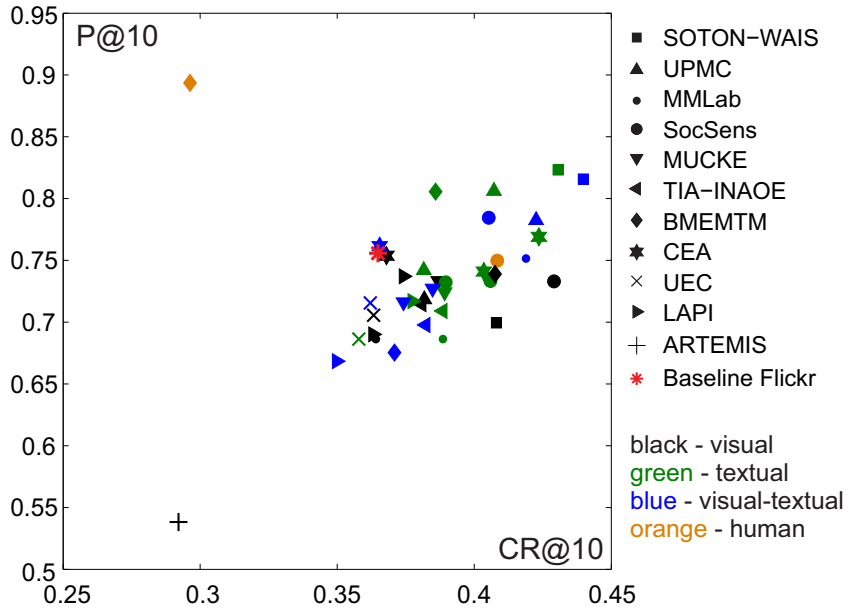

Fig. 4: Precision vs. cluster recall averages at 10 images.

to remain more or less within the same limits. The explanation of this effect may come from the fact that results obtained with keywords are less accurate and tend to spread over a higher number of users, which has the potential of increasing the diversity.

Overall, retrieval using only keywords seems to be more efficient for a small number of images (i.e., less than 10), while the inclusion of the GPS information leads to better results for a higher number of images (see F1 measure in Table 2). However, there is no big difference between the two. This is due to the fact that the retrieval with GPS coordinates includes also the keywords. Another factor is that in general few images are provided with the GPS information (e.g., approximately 60% of the images from *testset* do not have GPS information [21]).

Given the fact that GPS information is not always available, in the following the focus will be put on analyzing the overall results achieved over the entire *testset* collection.

5.2 Evaluation per modality

In this section the focus is on analyzing the influence of the modality on the diversification performance. Figure 4 plots overall precision against recall averages for all participant runs at a cutoff at 10 images. Each modality, i.e.,

visual, text, visual-text and human-based analysis, is depicted with different colors.

Concerning the *visual approaches*, highest diversification is achieved with a Greedy optimization of VLAD-SURF (Vector of Locally Aggregated Descriptors - Speeded Up Robust Features) descriptors, CR@10= 0.4291 — SocSens run1 [25] (see Section 4). The authors employed an optimized version of VLAD-SURF representations [42] that include multiple vocabulary aggregations, joint dimensionality reduction with PCA and whitening. On the other end, lowest diversification is provided by a matching graph approach also with feature point information, namely Hessian-Affine co-variant regions [45] along with the RootSIFT (Scale Invariant Feature Transform) descriptors [46], CR@10=0.2921 — ARTEMIS [30]. However, results show that using simple color information like color histograms and detection of face regions can still achieve similar recall ratios, e.g., CR@10=0.4076 — BMEMTM run1 [24]. Therefore, differences between local, such as SIFTs, and global descriptors, such as color, are not inherently very large. The difference in performance is mainly related to the method and to the way the descriptors are integrated.

Compared to visual, *text-based approaches* tend to provide better results (see the green points distribution in Figure 4). Highest diversification is achieved using a re-ranking with the Lucene engine and Greedy Min-Max optimization, CR@10=0.4306 — SOTON-WAIS run2 [20]. Data were represented with time related information: time user — images taken by the same user within a short time period are likely to be similar; and month delta — images have increasing similarity with closer month of year. Lowest diversification is achieved with a classic bag-of-words of TF-IDF data and web inspired ranking, namely CR@10=0.3579 — UEC run2 [28].

Surprisingly, *human-based approaches* were less effective than the automatic ones as users tend to maximize precision at the cost of diversity, e.g., BMEMTM run4 [24] (see Section 4) achieves P@10=0.8936 but CR@10 is only 0.2963. This is also visible from the visual ranking experiment conducted in Section 5.6 that show how visually similar images are when selected solely by users (see an example in Figure 7). However, human-machine integration is able to provide improvement also for the diversity part, e.g., CR@10=0.4048 — SocSens run4 [25].

Overall, the best performing approach is a *multimodal* one (i.e., text-visual in our case). It allows to achieve a CR@10=0.4398 — SOTON-WAIS run3 [20] (multimodal integration is achieved by averaging individual descriptor image similarity matrices) — which represents an improvement over the diversification of the state-of-the-art Flickr initial results with more than one additional image class (see the red asterisk in Figure 4).

5.3 Ranking stability analysis

Table 3 presents the official ranking of the best team approaches for various cutoff points (highest values are represented in bold). Reported values are

Table 3: Precision and cluster recall averages for best team runs (ranking according to the CR@10 official metrics).

team best run	*metrics*	@10	@20	@30	@40	@50
SOTON-WAIS run3 [20]	P	**0.8158**	**0.7788**	0.7414	0.7059	0.6662
	CR	**0.4398**	0.6197	0.7216	0.7844	0.8243
	F1	**0.5455**	**0.6607**	0.7019	0.7117	0.7037
SocSens run1 [25]	P	0.733	0.7487	**0.7603**	0.7145	0.5915
	CR	0.4291	**0.6314**	0.7228	0.7473	0.7484
	F1	0.5209	0.6595	0.7087	0.6922	0.6259
CEA run2 [29]	P	0.769	0.7639	0.7565	**0.7409**	**0.7153**
	CR	0.4236	0.6249	0.7346	0.8148	**0.8668**
	F1	0.5227	0.6593	**0.7158**	**0.7448**	**0.7508**
UPMC run3 [21]	P	0.7825	0.73	0.7254	0.7099	0.6891
	CR	0.4226	0.6268	0.747	0.8154	0.854
	F1	0.53	0.6498	0.7078	0.7301	0.7308
MMLab run3 [26]	P	0.7515	0.7404	0.7335	0.7185	0.697
	CR	0.4189	0.6236	**0.7492**	**0.8205**	0.8653
	F1	0.5174	0.6514	0.7114	0.735	0.7386
BMEMTM run1 [24]	P	0.7389	0.7164	0.7182	0.7115	0.6927
	CR	0.4076	0.6139	0.7184	0.7935	0.844
	F1	0.5066	0.6363	0.6908	0.7204	0.7284
MUCKE run2 [27]	P	0.7243	0.7228	0.7183	0.708	0.6884
	CR	0.3892	0.5749	0.6877	0.7684	0.8306
	F1	0.4905	0.6182	0.679	0.7106	0.7232
TIA-INAOE run2 [22]	P	0.7091	0.7136	0.7146	0.7045	0.6851
	CR	0.3885	0.5732	0.6897	0.7719	0.8228
	F1	0.4801	0.6102	0.6744	0.706	0.714
LAPI run2 [23]	P	0.717	0.7111	0.6896	0.6477	0.5795
	CR	0.3774	0.5734	0.682	0.7472	0.7722
	F1	0.4736	0.6078	0.6579	0.6644	0.6322
baseline Flickr	P	0.7558	0.7289	0.7194	0.708	0.6877
	CR	0.3649	0.5346	0.6558	0.7411	0.7988
	F1	0.4693	0.5889	0.6576	0.6938	0.7065
UEC run1 [28]	P	0.7056	0.7092	0.7076	0.6948	0.6752
	CR	0.3633	0.5448	0.6743	0.7572	0.8154
	F1	0.4617	0.5926	0.6618	0.6936	0.7068
ARTEMIS run1 [30]	P	0.5383	0.3379	0.2269	0.1702	0.1361
	CR	0.2921	0.3306	0.331	0.331	0.331
	F1	0.3653	0.3194	0.2578	0.216	0.186

averages over all the locations in the data set. In addition to the information from Figure 4, Table 3 reveals in the first place the fact that the precision tends to decrease with the higher precision cut-offs. This is motivated by the fact that increasing the number of results also increases the probability of including non-relevant pictures as in general the best matches tend to accumulate among the first returned results. Second, in contrast to the precision, cluster recall and thus diversity, increases with the number of pictures. This result is intuitive as the more pictures we retrieve, the more likely is to include a representative picture from each of the annotated categories.

Table 4: Ranking stability analysis for subsets of the dataset of different sizes.

Subset size	*metrics*	10	50	**100**	150	200	250	300
Spearman's ρ	CR@10	0.61	0.86	**0.93**	0.96	0.97	0.98	0.99
	P@10	0.74	0.92	**0.96**	0.98	0.99	0.99	0.99
	F1@10	0.64	0.89	**0.95**	0.97	0.98	0.99	0.99
Kendall's τ	CR@10	0.45	0.70	**0.79**	0.84	0.88	0.91	0.94
	P@10	0.59	0.79	**0.86**	0.9	0.93	0.95	0.97
	F1@10	0.48	0.74	**0.82**	0.87	0.91	0.93	0.95

To determine the statistical significance of the results and thus to examine the relevance of the dataset, a stability test was run [9]. Stability is examined by varying the number of topics which are used to compute performance. Stability tests are run with different topic subset sizes, which are compared to the results obtained with the full test set (346 topics). For each topic subset, 100 random topic samplings are performed to obtain stable averages. Spearman's rank correlation [40] (ρ — a measure of statistical dependence) and Kendall's tau coefficients [41] (τ — a measure of the association between two measured quantities) are used to compare the obtained CR@10, P@10 and F1@10 performances. The results for different subset sizes are presented in Table 4.

The results confirm the intuition that the more topics are evaluated, the more stable the rankings are. The values of both coefficients increase with the number of topics, with a faster pace for Spearman's ρ compared to Kendall's τ. The correlation is weaker for CR@10 compared to P@10 and F1@10 values naturally fall between CR@10 and P@10.

More importantly, the results indicate that little change in run ranking appears when at least 100 topics are used. Strong correlations for both Spearman's ρ and Kendall's τ are obtained starting from this point. For instance, with 100 topics, the first coefficient reaches between 0.93, 0.96 and 0.95 values for CR@10, P@10 and F1@10 while the corresponding values for Kendall's τ are 0.79, 0.86 and 0.82. To interpret Kendall's τ [10], the coefficient can be used to compute the ratio between concordant and discordant pairs in the two sets using:

$$r = \frac{1+\tau}{1-\tau} \tag{4}$$

For instance, at $\tau = 0.82$, obtained for F1@10 with 100 topics, there are 10 times more concordant pairs than discordant pairs in the compared rankings.

The size of the test set is clearly sufficient to ensure statistical stability of the ranking and therefore of the results. In the future, it might even be possible to reduce its size with very little loss of ranking stability.

[10] `http://www.rsscse-edu.org.uk/tsj/bts/noether/text.html`

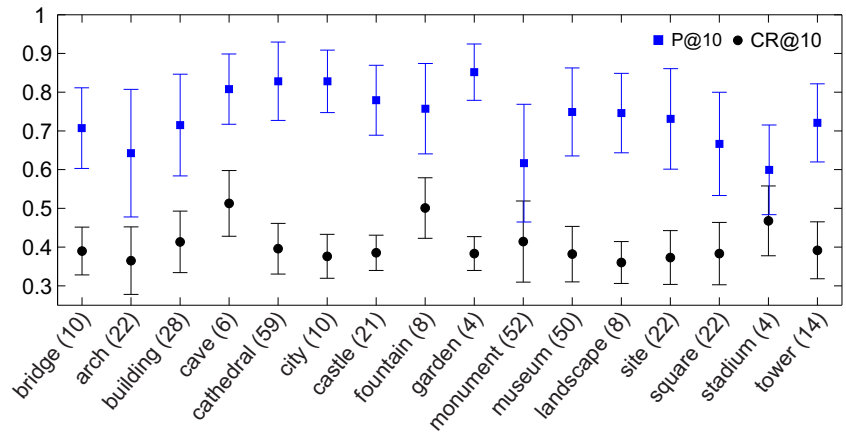

Fig. 5: Precision and cluster recall averages for best team runs per location type at 10 images (standard deviation is depicted with vertical lines). Numbers in the brackets represent the number of locations.

5.4 Evaluation per topic class

Depending on their nature, some landmarks are more complex than others in terms of visual characteristics. For instance, a monument can result in a limited number of perspectives that may capture views from different angles at different times of the day or the year, a cathedral involves in addition to that inside views or shots of the artifacts while an archeological site is inherently more diversified due to its higher geographic spread. Consequently it is worth investigating the influence of the location type on the diversification performance.

Figure 5 presents average best team runs in terms of cluster recall and precision on a per location basis. All categories were used as given by Wikipedia. While on average most of the locations seem to provide similar diversification possibilities, in particular highest diversification is achieved for caves, fountains and monuments. In terms of precision, very accurate predictions can be made for locations such as cathedrals or gardens which reach a precision above 0.9. The highest variability of the results is reported for arch and monuments which may be due to their intra-class variability (e.g., there are many types of monuments).

These results show that depending on the method and location characteristics, very accurate diversification is achievable, e.g., for caves the highest performance in terms of cluster recall is up to 0.6 (at 10 images, as previously mentioned, the highest possible value is 0.77, see beginning of Section 5) while precision is up to 0.9. Similar results are achieved for fountains.

Table 5: Expert vs. crowd annotations — precision and cluster recall averages for team best runs (selection made on expert ground truth).

team best run	*expert GT*			*crowd GT*		
	P@10	CR@10	F1@10	P@10	CR@10	F1@10
SOTON-WAIS run3 [20]	**0.8755**	0.4129	0.5403	**0.7714**	0.745	0.7301
SocSens run1 [25]	0.7959	0.4139	0.5314	0.7286	0.7636	0.7235
CEA run2 [29]	0.8265	0.4081	0.5242	0.7082	0.7287	0.6835
UPMC run3 [21]	0.8408	**0.4151**	**0.5441**	0.749	**0.788**	**0.7421**
MMLab run3 [26]	0.8347	0.4086	0.5376	0.6939	0.7569	0.6887
BMEMTM run1 [24]	0.8286	0.3997	0.5292	0.6857	0.7302	0.6722
MUCKE run2 [27]	0.8163	0.3716	0.5019	0.7204	0.7406	0.6997
TIA-INAOE run2 [22]	0.7837	0.3539	0.4783	0.6755	0.7258	0.6639
LAPI run2 [23]	0.8224	0.3920	0.5140	0.7163	0.7407	0.6941
baseline Flickr	0.7980	0.3345	0.4558	0.6816	0.6643	0.6269
UEC run1 [28]	0.7857	0.3755	0.4927	0.6959	0.7198	0.6657
ARTEMIS run1 [30]	0.6857	0.3453	0.4483	0.6449	0.7510	0.6615

5.5 Expert vs. crowd annotations

Performance assessment depends on the subjectivity of the ground truth, especially for the diversification part. The following experiment consists of comparing both results achieved with expert and crowd annotations. Table 5 presents the best team runs (highest results are depicted in bold) determined on a selection of 50 locations from the *testset* that was annotated also using crowd-workers, see Section 3.2.3. For each location, 3 different crowd generated diversity annotations were retained and we report the average metrics.

Although precision remains more or less similar in both cases, cluster recall is significantly higher for the crowd annotations. This is mainly due to the fact that workers tend to under-cluster the images for time reasons. Also, different than the experts, crowd-workers were not familiar with the location characteristics and this introduces a certain amount of variability in their annotations (see also Table 1).

Interestingly, regardless of the ground truth, the improvement in diversity of the baseline is basically the same: 0.0784 for experts compared to 0.0807 for the crowd, which shows that results are simply translated but the relevance is still comparable.

Crowd annotations are an attractive alternative to expert annotations, being fast — in the order of hours compared to expert ones that require weeks — while the performance may provide satisfactory results.

5.6 Human visual ranking

All previous result evaluations were carried out by computing objective numeric measures such as the recall and precision metrics. It is interesting to see however how the results are also perceived by the common user, which in the end is "the consumer of the retrieval system".

Table 6: Human-based visual ranking — best ranked team runs.

"Asinelli Tower"		*"Arc de Triomf"*	
best team run	average score	best team run	average score
SOTON-WAIS run2 [20]	1.67	SocSens run1 [25]	1.33
LAPI run1 [23]	4	TIA-INAOE run2 [22]	3.67
SocSens run1 [25]	4	SOTON-WAIS run2 [20]	5.33
UEC run3 [28]	8.67	CEA run5 [29]	5.67
UPMC run2 [21]	8.67	LAPI run1 [23]	7.67
BMEMTM run3 [24]	9.33	MMLab run1 [26]	10
MMLab run3 [26]	9.33	MUCKE run5 [27]	10.67
TIA-INAOE run2 [22]	12.33	UPMC run1 [21]	12.33
MUCKE run5 [27]	14.33	BMEMTM run1 [24]	13.67
CEA run3 [29]	18	UEC run1 [28]	23

The final experiment consists of a subjective evaluation. For each of the submitted system runs, the 10 highest-ranked images (used for official ranking) were printed on separate sheets of paper. Then, prints were presented to human observers who were asked to rank them according to their own perception of diversity and relevance (the definition in Section 3.2 was adopted). Each run was assigned a score ranging from 1 to the number of runs (38) — lowest numbers correspond to the highly diversified and visually appealing runs. It should be noted that there is no perfect correlation between image relevance and aesthetic appeal, therefore runs with relevant pictures may differ one another from the visual quality point of view. Prior to the test, observers were made familiar with the location characteristics to be able to determine which images are relevant or not. The task was not time restricted.

For experimentation we selected two of the locations from the *testset*, namely "Asinelli Tower" in Italy that in general provides a high diversity of the retrieved pictures but with variable relevance; and "Arc de Triomf" in Spain which has in general a high relevance of the pictures but comes with low diversity. The experiment was conducted with 3 observers (2 males and one female). Final ranking of the runs was determined after averaging all the observers' individual scores.

The results are summarized in Table 6 by presenting each team's best visual run. What is interesting to notice is that the best performing systems in terms of objective evaluation are also the systems that provide the highest ranked runs in term of subjective visual evaluation, i.e., SOTON-WAIS [20] and SocSens [25]. Both systems use Greedy optimization to ensure the diversification of the results.

To have a measure of the agreement between human visual rankings, we use again the Spearman's rank correlation coefficient [40]. We selected this measure in contrast to Cohen's kappa based statistics, because it is better suited for comparing different system rankings, as it is our case. The obtained results are presented in Table 7 and they show that correlations are moderate for *"Asinelli Tower"* and strong for *"Arc de Triomf"* (for reference, a value

Table 7: Spearman's rank correlation coefficients of the human rankings.

	"Asinelli Tower"			*"Arc de Triomf"*		
Observer	1	2	3	1	2	3
1 (male)	1	0.617	0.555	1	0.833	0.695
2 (male)	-	1	0.583	-	1	0.733
3 (female)	-	-	1	-	-	1

between .40 and .59 corresponds to a "moderate" correlation, between .60 and .79 to a "strong" correlation and between .80 and 1.0 to a "very strong" correlation[11]). This result reflects well the difference in visual complexity of the two locations analyzed here, with the simpler one obtaining higher correlations than the other one.

Finally, in Figures 6 and 7 we illustrate for comparison the images provided by the initial Flickr results, the best visual systems and the lowest ranked systems for the two locations. For "Asinelli Tower" that in general comes with a high diversity of the images the limitation of the lowest ranked run comes from the high number of non-relevant pictures provided and also from the number of pictures that do not show the main characteristics of the target — CEA run5 [29], average rank 33 (CR@10=0.3077); whereas the highest ranked system is able to provide only relevant pictures — SOTON-WAIS run2 [20], average rank 1.67 (CR@10=0.3846). For "Arc de Triomf" where the diversity of the retrieved images is in general low, the highest ranked system is able to provide significant diversification with only one partially non-relevant picture (tagged this way due to the people in focus) — SocSens run1 [25], average rank 1.33 (CR@10=0.4615). What is interesting to see here is that the lowest ranked system is a human run — BMEMTM run4 [24], average rank 38 (CR@10=0.3077) — that outputs only relevant pictures but with the cost of their diversity, as almost all depict quasi-similar characteristics of the location. One explanation is the fact that human observers are tempted to select images that are close the most representatives (common) picture of the location which displays the monument from the front. This holds also for the "Asinelli Tower" location where the human-based run was the penultimate ranked run with an average score of 32.67 (not displayed here for brevity).

6 Conclusions and Outlook

This article introduces a benchmarking framework for results diversification of social image retrieval and describes the related task run in the MediaEval 2013 campaign. The strong participation in a first year (24 teams registered and 11 crossed the finish line) shows the strong interest of the research community in the topic. Similar to the strong impact of other evaluation campaigns in

[11] for the interpretation of Spearman's rank correlation coefficient values, see `http://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf`

(a) baseline Flickr initial results

(b) highest visually ranked run — SOTON-WAIS run2 [20]

(c) lowest visually ranked run — CEA run5 [29]

Fig. 6: Visual comparison of the results for "Asinelli Tower" (Italy). Flickr image credits (from left to right and top to bottom): (a) lorkatj, leonardo4it, kyle NRW, Viaggiatore Fantasma, kyle NRW, sara zollino, Alessandro Capotondi, magro_kr (2 images), Funchye; (b) adrian, acediscovery, kondrag, lorenzaccio*, sara zollino, pietroizzo, greenblackberries, Xmansti, Argenberg (2 images), pietroizzo; (c) lorkatj, leonardo4it, kyle NRW, Viaggiatore Fantasma, sara zollino, Alessandro Capotondi, magro_kr (2 images), Funchye, Dimitry B. Non-relevant images are marked with a red X (according to definition in Section 3.2). Only the first 10 ranks are displayed (official cutoff point).

(a) baseline Flickr initial results

(b) highest visually ranked run — SocSens run1 [25]

(c) lowest visually ranked run — BMEMTM run4 [24]

Fig. 7: Visual comparison of the results for "Arc de Triomf" (Spain). Flickr image credits (from left to right and top to bottom): (a) Sam Kelly, tetegil, euthman, Peter Zoon, 3nt, myBCN - Barcelona Expert (last 5 images); (b) Bilbopolit, urgetopunt, myBCN - Barcelona Expert, Mark & Gideon, fpeault, Kansas Sebastian, mazlov, leoglenn_g, urgetopunt, . SantiMB .; (c) karolajnat, mr-numb, craggyisland21, . SantiMB ., sincretic, leoglenn_g, tetegil, pbb, Andy_Mitchell_UK, teachandlearn. Un-relevant images are marked with a red X (according to definition in Section 3.2). Only the first 10 ranks are displayed (official cutoff point).

multimedia retrieval [2,3,49] an important impact can also be expected from this task as already the analysis of this paper shows. Several groups increased specific aspects of the results on the strong Flickr baseline, particularly linked to diversity. Approaches combining a large variety of modalities from manual re-ranking, GPS to visual and text attributes have the potential to improve results quality and adapt to what users may really want to obtain as results, which can be situation-dependent. Detecting objects such as faces was also used. Via the analysis of the clusters of relevant images, several categories can likely be deduced and used in connection with detectors for these aspects to optimize results.

The crowdsourcing part of the relevance judgments is clearly an option as the results described in the paper show. There are differences in the results but the effort to cost ratio is an important part and crowdsourcing can likely help to create much larger resources with a limited funding. Strict quality control seems necessary to assure the crowdsources quality and this can likely also help to obtain better results in the future.

For a continuation of the evaluation campaign it seems important to look into criteria that can stronger discriminate the runs, so basically making the task harder. More clusters are an option, or a hierarchy of clusters. A larger collection is also an option but creating diversity ground truth for large collections is tedious and expensive. Crowdsourcing could be a valid approach also for this, as the experiments show.

Overall, results are stable with the number of test topics chosen. This number could even be reduced with little negative effect on stability. Several outcomes of analyzing the runs of the participants show that multimodal approaches often perform best. Greedy optimization seems to work well providing some of the highest quality results. Manual approaches tend to favor relevance over diversity. The analysis outlined in this paper gives several clear ideas on how to obtain better results and how to optimize results for both diversity and precision. This can likely lead to several other applications that can show their optimized performance on this publicly available resource.

7 Acknowledgments

This work was supported by the following projects: CUbRIK (`http://www.cubrikproject.eu/`), PROMISE (`http://www.promise-noe.eu/`) and MUCKE (`http://ifs.tuwien.ac.at/~mucke/`). We acknowledge also the MediaEval Benchmarking Initiative for Multimedia Evaluation (`http://www.multimediaeval.org/`).

References

1. A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, "Content-based image retrieval at the end of the early years", IEEE Transaction on Pattern Analysis and Machine Intelligence, 22(12), pp. 1349 - 1380, 2000.

2. P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A.F. Smeaton, G. Quéenot, TRECVID 2013 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics, Proceedings of TRECVID 2013, http://www-nlpir.nist.gov/projects/tvpubs/tv13.papers/tv13overview.pdf, NIST, USA, 2013.
3. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.
4. R. Datta, D. Joshi, J. Li, J.Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age", ACM Comput. Surv., 40(2), pp. 1-60, 2008.
5. R. Priyatharshini, S. Chitrakala, "Association Based Image Retrieval: A Survey", Mobile Communication and Power Engineering, Springer Communications in Computer and Information Science, 296, pp 17-26, 2013.
6. L. McGinty, B. Smyth, "On the role of diversity in conversational recommender systems", International Conference on Case-Based Reasoning, pp. 276-290, 2003.
7. 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.
8. M.L. Paramita, M. Sanderson, P. Clough, "Diversity in Photo Retrieval: Overview of the ImageCLEF Photo Task 2009", ImageCLEF 2009.
9. T. Tsikrika, J. Kludas, A. Popescu, "Building Reliable and Reusable Test Collections for Image Retrieval: The Wikipedia Task at ImageCLEF", IEEE Multimedia, 19(3), pp. 24-33, 2012.
10. 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.
11. 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.
12. 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, LNCS 8325, pp. 25-36, 2014.
13. R. Agrawal, S. Gollapudi, A. Halverson, S. Ieong, "Diversifying search results", ACM International Conference on Web Search and Data Mining, Barcelona, Spain, 2009.
14. X. Zhu, A. Goldberg, J. V. Gael and D. Andrzejewski, "Improving Diversity in Ranking using Absorbing Random Walks", Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2007.
15. E. Vee, U. Srivastava, J. Shanmugasundaram, P. Bhat, S.A. Yahia, "Efficient computation of diverse query results", IEEE International Conference on Data Engineering, pp. 228 - 236, 2008.
16. Z. Huang, B. Hu, H. Cheng, H. Shen, H. Liu and X. Zhou, "Mining near-duplicate graph for cluster-based reranking of web video search results", ACM Transactions on Information Systems, vol. 28, pages 22:1-22:27, USA, November 2010.
17. M.R. Vieira, H.L. Razente, M.C.N. Barioni, M. Hadjieleftheriou, D. Srivastava, C. Traina Jr., V.J. Tsotras, "On Query Result Diversification", IEEE International Conference on Data Engineering, pp. 1163 - 1174, 11-16 April, Hannover, Germany, 2011.
18. Dang, Van and Croft, W. Bruce, "Diversity by Proportionality: An Election-based Approach to Search Result Diversification", ACM International Conference on Research and Development in Information Retrieval, pp. 65-74, Portland, Oregon, USA, 2012.
19. MediaEval 2013 Workshop, Eds. M. Larson, X. Anguera, T. Reuter, G.J.F. Jones, B. Ionescu, M. Schedl, T. Piatrik, C. Hauff, M. Soleymani, co-located with ACM Multimedia, Barcelona, Spain, October 18-19, CEUR-WS.org, ISSN 1613-0073, Vol. 1043, http://ceur-ws.org/Vol-1043/, 2013.
20. N. Jain, J. Hare, S. Samangooei, J. Preston, J. Davies, D. Dupplaw, P. Lewis, "Experiments in Diversifying Flickr Result Sets", Working Notes Proceedings [19], 2013.
21. C. Kuoman, S. Tollari, M. Detyniecki, "UPMC at MediaEval 2013: Relevance by Text and Diversity by Visual Clustering", Working Notes Proceedings [19], 2013.
22. H.J. Escalante, A. Morales-Reyes, "TIA-INAOE's Approach for the 2013 Retrieving Diverse Social Images Task", Working Notes Proceedings [19], 2013.

23. A-L. Radu, B. Boteanu, O. Pleş, B. Ionescu, "LAPI @ Retrieving Diverse Social Images Task 2013: Qualitative Photo Retrieval using Multimedia Content", Working Notes Proceedings [19], 2013.
24. G. Szűcs, Z. Paróczi, D.M. Vincz, "BMEMTM at MediaEval 2013 Retrieving Diverse Social Images Task: Analysis of Text and Visual Information", Working Notes Proceedings [19], 2013.
25. D. Corney, C. Martin, A. Göker, E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, L. Aiello, B. Thomee, "SocialSensor: Finding Diverse Images at MediaEval 2013", Working Notes Proceedings [19], 2013.
26. B. Vandersmissen, A. Tomar, F. Godin, W. De Neve, R. Van de Walle, "Ghent University-iMinds at MediaEval 2013 Diverse Images: Relevance-Based Hierarchical Clustering", Working Notes Proceedings [19], 2013.
27. A. Armagan, A. Popescu, P. Duygulu, "MUCKE Participation at Retrieving Diverse Social Images Task of MediaEval 2013", Working Notes Proceedings [19], 2013.
28. K. Yanai, D.H. Nga, "UEC, Tokyo at MediaEval 2013 Retrieving Diverse Social Images Task", Working Notes Proceedings [19], 2013.
29. A. Popescu, "CEA LISTs Participation at the MediaEval 2013 Retrieving Diverse Social Images Task", Working Notes Proceedings [19], 2013.
30. A. Bursuc, T. Zaharia, "ARTEMIS @ MediaEval 2013: A Content-Based Image Clustering Method for Public Image Repositories", Working Notes Proceedings [19], 2013.
31. B. Ionescu, M. Menéndez, H. Müller, A. Popescu, "Retrieving Diverse Social Images at MediaEval 2013: Objectives, Dataset and Evaluation", Working Notes Proceedings [19], 2013.
32. B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, B. Loni, "Div400: A Social Image Retrieval Result Diversification Dataset", ACM Multimedia Systems, 19-21 March, Singapore, 2014.
33. B. Ionescu, A. Popescu, H. Müller, M. Menéndez, A.-L. Radu, "Benchmarking Result Diversification In Social Image Retrieval", IEEE International Conference on Image Processing, October 27-30, Paris, France 2014.
34. L. Ballan, M. Bertini, T. Uricchio, A. Del Bimbo, "Data-driven approaches for social image and video tagging", Multimedia Tools and Applications, DOI 10.1007/s11042-014-1976-4, 2014.
35. X. Li, C.G.M. Snoek, M. Worring, "Learning Social Tag Relevance by Neighbor Voting", IEEE Transactions on Multimedia, 11(7), pp. 1310 - 1322, 2009.
36. A. Popescu, G. Grefenstette, "Social Media Driven Image Retrieval", ACM ICMR, April 17-20, Trento, Italy, 2011.
37. J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit", Psychological Bulletin, Vol. 70(4), pp. 213-220, 1968.
38. J. Cohen, "A coefficient of agreement for nominal scales", Educational and Psychological Studies Measurement, Vol. XX(1), pp. 37-46, 1960.
39. J.J. Randolph, "Free-Marginal Multirater Kappa (multirater κfree): an Alternative to Fleiss Fixed-Marginal Multirater Kappa", Joensuu Learning and Instruction Symposium, 2005.
40. A. Lehman, "Jmp For Basic Univariate And Multivariate Statistics: A Step-by-step Guide", Cary, NC: SAS Press. p. 123. ISBN 1-59047-576-3, 2005.
41. D. Wilkie, "Pictorial Representation of Kendall's, Rank Correlation Coefficient", Teaching Statistics 2, pp. 76-78, 1980.
42. E. Spyromitros-Xioufis, S. Papadopoulos, I. Kompatsiaris, G. Tsoumakas, I. Vlahavas, "An empirical study on the combination of SURF features with VLAD vectors for image search", International Workshop on Image Analysis for Multimedia Interactive Services, Dublin, Ireland, 2012.
43. Y. Jing, S. Baluja, "Visualrank: Applying pagerank to large-scale image search", IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(11):1870-1890, 2008.
44. X.-Q. Cheng, P. Du, J. Guo, X. Zhu, Y. Chen, "Ranking on data manifold with sink points", IEEE Transactions on Knowledge and Data Engineering, 25(1):177-191, 2013.
45. Perdoch, M., Chum, O., Matas, J., "Efficient Representation of Local Geometry for Large Scale Object Retrieval", IEEE Conf. on Computer Vision and Pattern Recognition, 2009.

46. Arandjelovic, R., Zisserman, A., "Three things everyone should know to improve object retrieval", IEEE Conf. on Computer Vision and Pattern Recognition, 2012.
47. T. Deselaers, T. Gass, P. Dreuw, H. Ney, "Jointly Optimising Relevance and Diversity in Image Retrieval", ACM Int. Conf. on Image and Video Retrieval, 2009.
48. A. García Seco de Herrera, J. Kalpathy-Cramer, D. Demner Fushman, S. Antani, H. Müller, Overview of the ImageCLEF 2013 medical tasks, Working Notes of CLEF 2013 (Cross Language Evaluation Forum), Valencia, Spain, 2013.
49. T. Tsikrika, Theodora, A. García Seco de Herrera, H. Müller, Assessing the Scholarly Impact of ImageCLEF, Springer Lecture Notes in Computer Science (LNCS), pages 95-106, 2011.