Benchmarking Image Retrieval Diversification Techniques for Social Media

Bogdan Ionescu, *Senior Member, IEEE,* Maia Rohm, Bogdan Boteanu, Alexandru Lucian Gînscă, Mihai Lupu, and Henning Müller, *Member, IEEE,*

***Abstract*—Image retrieval has been an active research domain for over 30 years and historically it has focused primarily on precision as an evaluation criterion. Similar to text retrieval, where the number of indexed documents became large and many relevant documents exist, it is of high importance to highlight diversity in the search results to provide better results for the user. The Retrieving Diverse Social Images Task of the MediaEval benchmarking campaign has addressed exactly this challenge of retrieving diverse and relevant results for the past years, specifically in the social media context. Multimodal data (e.g., images, text) was made available to the participants including metadata assigned to the images, user IDs, and precomputed visual and text descriptors. Many teams have participated in the task over the years. The large number of publications employing the data and also citations of the overview articles underline the importance of this topic. In this paper, we introduce these publicly available data resources as well as the evaluation framework, and provide an in-depth analysis of the crucial aspects of social image search diversification, such as the capabilities and the evolution of existing systems. These evaluation resources will help researchers for the coming years in analyzing aspects of multimodal image retrieval and diversity of the search results.**

I. Introduction

IMAGE retrieval has been an extremely active research domain over the past 30 years [1], [2]. Starting with text-based retrieval of images and then moving towards content-based image retrieval and multimodal approaches, the techniques have constantly evolved to high quality of retrieval and increasingly large data sets [3], [4]. The evaluation of retrieval approaches has traditionally focused on early precision in retrieval results and on mean average precision (MAP) [5]–[7], and for specific applications, e.g., patent retrieval, on recall. With increasingly large data sets and many potentially relevant images, precision as an evaluation criterion is not sufficient anymore and requires complementary measures.

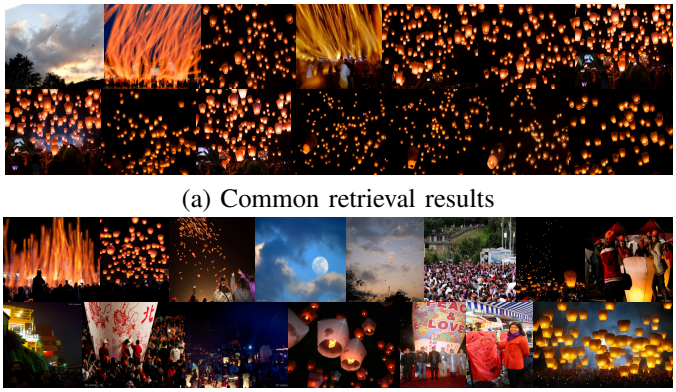
In most cases, systems aim to improve the relevance of the results assuming that the results for a query are single topic. This is not an accurate assumption anymore in the context of the current Internet, because many of the queries cover different aspects, i.e., sub-topics. For instance, objects in images show different information and have different contexts, landmarks can be captured in various conditions and angles, e.g., day-night, close-far, bicycles serve different usages conditions, e.g., city, mountain, road, vehicles are of different types, and so on. An effective retrieval system should also take into account the *diversification* of the results [8]. An example is provided in Figure 1.

To improve the diversity of search results, one has to consider the multiple and diverse topics, contexts, intents, and interpretations of a certain query. Increasing the diversity increases also the efficiency and usefulness of the system via providing a wider selection of results and therefore, a higher chance that they address the user real needs. A concrete example are the recommender systems, where the users' satisfaction increases with the diversification of the results. With this concept, cluster recall was introduced as a measure for diversity in image retrieval [8]. The Retrieving Diverse Social Images Task, we are introducing in this paper, has been organized under the MediaEval Benchmarking Initiative for Multimedia Evaluation and has evaluated such approaches over the past years [9]–[13].

Another important aspect of image retrieval is the availability of many input sources, e.g., not only features that represent the visual image content and textual metadata, but also information on the person posting data, tags added by other persons and possible GPS (Global Positioning System) data. Some of this information may represent what is in the image, others what the image is about but also emotional responses, for example the feeling that an image evokes and the context in which it was taken. The Retrieving Diverse Social Images Task addressed these aspects and created a benchmark framework so that practitioners could choose from a large number of data sources. Additionally, various visual- and text-based content descriptors were made available to limit the entry requirements for systems [9], [12] while focusing on diversification.

The large number of publications employing the various data sets from the task and the increasing number of citations underline the importance and the high impact of the task. With the public availability of resources, we expect the impact and usage of the resources to increase strongly over the coming years, similar to other related benchmarking campaigns [14]. Thus, it seems important to analyze the benchmark and to describe the main challenges and lessons learned to foster further research in the context of image retrieval diversification.

The remainder of the article is organized as follows. Section II positions our work in the context of the state of the art and highlights its contribution. Section III introduces the proposed social image search diversification benchmark framework: data sets, pre-computed content descriptors, available annotations, and evaluation methodology. Section IV investigates several crucial aspects in the context of image search diversification, such as the capabilities and the evolution of existing systems and presents experimental results. Section V concludes the paper and discusses future challenges.



(a) Common retrieval results

(b) Results after diversification

Fig. 1. Example of retrieval and diversification results for query "Pingxi Sky Lantern Festival" (results are truncated for visualization): (a) Flickr initial retrieval results; (b) diversification achieved with the approach from TUW [15] (best approach at MediaEval 2015).

II. Previous Work

Diversification is an actively researched topic in various domains ranging from web search and query result diversification [9], [16]–[19] to recommender systems [20]–[22] and summarization [23]–[25]. With the emerging availability of publicly available images, the importance of diversification of image data in the results is steadily growing. Following the structure of the current paper, we will first present previous work in benchmark creation, followed by a few reference articles in methods for diversification.

To this date, several benchmark initiatives promoted the development and comparability of approaches in the context of diversification: the ImageCLEF Photo Task (2008-2009) [8], [26], ImageCLEF Lifelogging Task (2017-2019) [27], TREC Web Track: Diversity Task (2009-2012) [28], and MediaEval Retrieving Diverse Social Images Task (2013-2017)[1]. In its last year, the *ImageCLEFPhoto Task* 2009 focused on image retrieval and diversity in a large collection of 498,000 press photos. There are 50 query topics and via pooling, a subset of the results were judged for relevance [8]. The quality of both images and annotations differs strongly from those of social media. Moreover, the tasks on relevance estimation and diversification are additionally supported by the availability of example images for all queries and cluster information for half of the queries. As the name says, the *ImageCLEF Lifelogging Task* (2017-2019) focuses on the retrieval of lifelog data according to predefined moments/events of everyday-life, such as certain activities, locations, or day-times. Although the task considers relevance and diversity as core aspects of the desired summarization, lifelog data differ significantly from social media data in terms of, e.g., data quality and variety in addition to missing user-provided metadata. The *TREC Web Track* (2009-2012) is considered a diversity task in the context of web documents only [28]. In contrast, the *MediaEval Retrieving Diverse Social Images Task* addresses the diversification of image search results with an explicit emphasis on the actual social media context. In its first years, 2013-2014, the task primarily focused on location-oriented queries. Starting from 2015, the task moved towards general-purpose and multi-concept queries. While this is a common and realistic scenario for the general user, it fosters the development of more general approaches since no reliable assumptions about the queries or underlying data can be made.

Because the existing diversity-focused benchmarks do not properly cover social media tasks, recent work in the context of a general query retrieval diversification of images from social media are commonly using self-collected data sets from Flickr [29]–[34]. For example, van Leuken et al. [29] collect Flickr data using 25 textually ambiguous and 50 textually unambiguous queries. However, no additional information on the data set is provided. Wang et al. [30] report experimental results on a Flickr data set of 104,000 images, collected using tag-based search for 52 distinct tags (e.g., *airshow*, *apple*, *beach*). Negi et al. [33] perform experiments on a small Flickr data set collected using 10 queries (e.g., *scorpion*, *jaguar*, *eagle*), with a total of 250 relevant images per query. Common core limitations of experiments on self-collected data concern the interpretability of the results due to often missing details on the data acquisition process and the limited comparability between related approaches. Moreover, self-collected data is commonly acquired using only single term queries. While this is a good starting point for research, real-world user queries commonly contain multiple concepts allowing to express their intent and information need more precisely (e.g., search for *graffiti on a wall* rather than only *graffiti*).

In terms of diversification methods, a broad range of articles addressed the diversification of images in the context of geographical applications (e.g., landmarks and general geographical summaries) [25], [35]–[39]. For example, Rudinac et al. [25] creates visual summaries of geographic areas using user-contributed images and related metadata. The approach is based on a Random Walk scheme with restarts over a graph that models relations between images, visual features, user-provided metadata, and the information on the uploader and commentators. Radu et al. [36] employed a crowd-sourcing approach to improve the initial results achieved by an automated visual analysis of the retrieval results for monument queries. To avoid the use of human expertise, Boteanu et al. [39] considered pseudo-relevance feedback, where user feedback is simulated by the selection of positive and negative examples from the initial query results. For a thorough survey of recent approaches, we refer to [9], [18]. Existing approaches in the context of geographical locations are not necessarily tailored to the characteristics of the application scenario (e.g., the availability of GPS information). However, location-based queries are usually well-defined and with a partially limited degree of visual diversity.

To explore the generalization ability of a given approach, thorough experiments on several data sets and application scenarios are required. A recent work in this direction is reported by Boato et al. [40]. The authors make use of visual saliency information for the diversification of image retrieval results and present experiments on two data sets: a self-combined collection of publicly available data sets in the context of object categorization and the Div150Cred

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

data set [12] addressing the diversification of POI (points of interest) images retrieved from Flickr. The reported results demonstrate a notable difference in the performance on the two application scenarios, i.e., while the improvement in the diversification on the object categories is significant, the difference on the location-based data set is marginal only. Images depicting different object categories are usually set around a centered main object in focus, which is in favor of the proposed approach. In contrast, location-based images (from social media) depict a higher degree of visual variation and do not necessarily follow common saliency rules. Another work demonstrating applicability across different application scenarios is presented by Desealers et al. [41]. The authors employ dynamic programming for the optimization of relevance and diversity in two scenarios: retrieval of natural images (ImageCLEF 2008 [42]) and retrieval of product images [41]. However, the authors provide only a qualitative evaluation on the product data, which limits the interpretability of the results. Additionally, the scenarios are build upon (semi-)professional photos and annotations, which differ notably from the characteristics of social media in terms of quality of both image data and associated metadata.

The current work brings additional value to the state of the art with the following main contributions: (i) It introduces a publicly available, common image search diversification benchmarking framework with explicit focus on social media aspects. The framework builds on current state-of-the-art retrieval technology (i.e., Flickr's relevance system), allowing to push diversification in priority. It comes with a large variety of data and query information (single, multi and adhoc topics) for complex scenarios; (ii) It provides an in-depth analysis of the crucial aspects of image search diversification, such as the capabilities and the evolution of existing systems thorough the analysis of the results from the MediaEval Retrieving Diverse Social Images 2013 — 2016 tasks. Experimental results highlight the social facets of the problem, the contribution of deep learning and user credibility information, the choice of feature combinations and fusion types, and various proposed approaches. To the best of our knowledge, this is the first comprehensive study covering all these core aspects of the social media diversification tasks.

III. Evaluation Framework

In this section, we present the components that make up the evaluation framework: the data sets (Section III-A), the content descriptions we provide (Section III-B), the ground truth (Section III-C), and the evaluation methodology that brings together all these data (Section III-D).

A. Data sets

Several data sets were designed and created for benchmarking image retrieval diversification capabilities with the explicit focus on the actual social media context. The social aspects are reflected both in the nature of the data (variable quality of photos and of metadata shared on social media, assessment of user tagging credibility, etc.) and in the methods employing the data. All the data consists of redistributable Creative Commons Flickr[2] and Wikipedia[3] data and are publicly released for reproducibility and comparability reasons [10]–[13]. Each of these data sets addresses different perspectives of the image diversification challenge and were validated during the annual MediaEval Benchmarking Initiative for Multimedia Evaluation as part of the Retrieving Diverse Social Images Task. An overview of the released data is presented in Table I. For three of the four data sets, *the concept of user tagging credibility* is used to provide additional input to the retrieval methods. Credibility, as the general concept covering trustworthiness and expertise but also quality and reliability is strongly debated in philosophy, psychology, and sociology. Automated credibility estimation is a recent trend in Web content analysis and is mostly applied to textual documents, such as tweets [43] or Web pages [44]. The presence of credibility related approaches in the multimedia domain is limited. Yamamoto and Tanaka [45] propose ImageAlert, a system that focuses on text-image credibility, while Benevenuto et al. [46] focus on the credibility of the users and aim to detect users distributing video spam rather than classifying the content itself. In Web 2.0 platforms, the quality of annotations provided by different users can vary strongly, e.g., Izadinia et al. [47] studied the tags of 269,642 Flickr images from the NUS-WIDE data set [48] for 81 manually labeled topics and observed that a tag has only a 62% chance of being correctly associated to images. The estimation of individual tag relevance is related to our view of user tagging credibility. The proposed estimates for user credibility are detailed in Section III-B3.

Div400 data set [10] focuses on the retrieval of photos form a predefined location (POI) in a tourism scenario, i.e., the user searches for a diversified photo summary for a target location. The data set consists of 396 landmark location queries (e.g., museums, monuments, bridges) ranging from very famous ones, e.g., "*Big Ben in London*", to less known, e.g., "*Palazzo delle Albere in Italy*". Each query location is provided with: location name, GPS coordinates (latitude and longitude), a link to its Wikipedia web page, a representative photo from Wikipedia, a ranked set of photos retrieved from Flickr with text and GPS queries using Flickr's default "relevance" algorithm (up to 150 photos), and their metadata (photo's id, title, description, tags, geotagging information, the date the photo was taken, owner's name, the number of times the photo has been viewed, the url link of the photo location from Flickr, license type, number of posted comments). The data set is divided into a development set (*devset*) containing 50 of the queries (5,118 Flickr photos) and a test set (*testset*) containing the remaining 346 queries (38,300 Flickr photos).

Div150Cred data set [12] uses the same use case as Div400, i.e., tourism, and is built on top of these data. It provides 300 queries with re-crawled information for ensuring up to 300 photos, and up to 5 representative images from Wikipedia for each query. It extends the metadata by providing also the userid from Flickr. The distribution into development and test set data is as follows: *devset* containing 30 queries

[2]https://www.flickr.com/
[3]https://en.wikipedia.org/

TABLE I
DATA SET STATISTICS (*devset* — DEVELOPMENT DATA, *testset* — TESTING DATA, *credibilityset* — DATA FOR ESTIMATING USER TAGGING CREDIBILITY, *single* ($_{st}$) — SINGLE TOPIC QUERIES, *multi* ($_{mt}$) — MULTI-TOPIC QUERIES, ++ — ENHANCED/UPDATED CONTENT, *POI* — LOCATION POINT OF INTEREST, *events* — EVENTS AND STATES ASSOCIATED WITH LOCATIONS, *general* — GENERAL PURPOSE AD HOC TOPICS).

	Div400 (2013)		**Div150Cred (2014)**			**Div150Multi (2015)**				**Div150Adhoc (2016)**		
	devset	*testset*	*devset*	*testset*	*credibilityset*	*devset*	*testset*$_{st}$	*testset*$_{mt}$	*credibilityset*	*devset*	*testset*	*credibilityset*
data source	2013	2013	2013++	2013++	2014	2014all	2015	2015	2014++	2015/2016	2016	2015++
#queries	50	346	30	123	300 POIs, 685 users, ∼3.6M + ∼12.3M (via *devset* & *testset*) image urls & metadata	153	69	70	300 POIs, 685 Flickr users, ∼3.6M + ∼27.1M (via *devset* & *testset*) image urls & metadata	70	64	300 POIs, 685 Flickr users, ∼3.6M image urls
content type	POI single	POI single	POI single	POI single		POI single	POI single	events multi		events/general multi	general multi	
#Wiki.img./query	1	1	1-5	1-5		1-5	1-5	-		-	-	
#images	5,118	38,300	8,923	36,452		45,375	20,700	20,694		20,757	18,717	
min #img./query	30	30	285	277		281	300	176		176	141	
avg. #img./query	102.4	110.7	297	296		297	300	296		297	292	
max #img./query	150	150	300	300		300	300	300		300	300	
descriptors	visual TF-IDF		visual, TF-IDF credibility			visual, TF-IDF credibility, CNN				TF-IDF & semantic vectors credibility, CNN		

(8,923 Flickr photos) and *testset* containing 123 queries (36,452 Flickr photos). The main contribution is however in the introduction of user tagging credibility information for diversification together with a dedicated data set (*credibilityset*). The proposed credibility information on user tagging attempts to provide an estimation of the global quality of tag-image content relationships for user's contributions [49], e.g., upload activity, coherence of the tag sets, and correspondence between tags and content of the images. This information is in particular valuable for exploiting the social context of the data. It gives an indication about which users are most likely to share representative images in Flickr. In this context, the *credibilityset* provides Flickr photo information (e.g., date of the photo, user's id, number of visualizations, GPS information) for about 300 locations and 685 different users (a total of 3.6M image links with metadata). Each user is assigned a manual credibility score (ground truth) which is determined as the average relevance score of all the user's photos (see Section III-C). Each user is also provided with the estimated credibility descriptors introduced in Section III-B3. User information is provided also for the users in *devset* and *testset* via a total of 12.3M image links with metadata. This data set is intended for training and designing user credibility related descriptors.

Div150Multi data set [11] uses the same tourist scenario as in the previous editions and increases the difficulty by addressing multi-topic queries about location specific events, location aspects, or general activities (e.g., "Oktoberfest in Munich", "Bucharest in winter"). The data set consists of information for 300 single- and multi-topic queries and each query is provided with up to 300 photos and up to 5 representative images from Wikipedia. In terms of metadata, we provide the same type of information as for Div150Cred. The Div150Multi development set (*devset*) consists of 153 queries (i.e., the complete Div150Cred data set [12], 45,375 Flickr photos). The user annotation credibility set (*credibilityset*) contains information for 300 locations and 685 users (the updated version of Div150Cred with 3.6M image links with metadata and 27.1M image links with metadata for users in *devset* and *testset*). Finally, the test set (*testset*) contains 139 queries: 69 one-concept location queries (20,700 Flickr photos) and 70 multi-concept queries related to events and states associated with locations (20,694 Flickr photos).

Div150Adhoc data set [13] addresses the diversification problem for a general ad-hoc image retrieval system, where general purpose multi-topic queries are used for retrieving the images (e.g., "animals at Zoo", "flying planes on blue sky", "hotel corridor"). The data set consists of information for 134 multi-topic queries and each query is provided with up to 300 photos. In terms of metadata, we provide the same type of information as for the previous data set, Div150Cred. Div150Adhoc provides a development set (*devset*) containing 70 queries (20,757 Flickr photos including 35 multi-topic queries related to events and states associated with locations from Div150Multi [11]), the user tagging credibility set (*credibilityset*) containing information for 300 location-based queries and 685 users (the updated version of Div150Multi, with 3.6M image links with metadata), a set providing semantic vectors for general English terms computed on top of the English Wikipedia (*wikiset*, see Section III-B2), for developing advanced text models, and a test set (*testset*) containing 65 queries (19,017 Flickr photos).

B. Content descriptors

To address a broader community, the data sets come with pre-computed content descriptors, namely:

1) General-purpose visual descriptors: (Div400, Div150Cred, Div150Multi) for each image, the following descriptors are provided which are known to perform well on image retrieval tasks: *global color naming histogram*: maps colors to 11 universal color names, i.e., "*black*", "*blue*", "*brown*", "*grey*", "*green*", "*orange*", "*pink*", "*purple*", "*red*", "*white*", and "*yellow*" [50]; *global Histogram of Oriented Gradients*: represents the HoG feature computed on 3 by 3 image regions [51]; *global color moments* on HSV (Hue-

Saturation-Value) color space: represents the first three central moments of an image color distribution: mean, standard deviation and skewness [52]; *global Locally Binary Patterns* on gray scale [53]; *global Color Structure Descriptor*: represents the MPEG-7 Color Structure Descriptor computed on the HMMD (Hue-Min-Max-Difference) color space [54]; *global statistics on gray level Run Length Matrix*: provides 11 statistics computed on gray level run-length matrices for 4 directions, e.g., Gray-Level Non-uniformity, High Gray-Level Run Emphasis, see [55]; and their *local spatial pyramid representations* (descriptors are computed on image blocks of 3 by 3 pixels and then merged into a global descriptor).

2) Text models and descriptors: (Div400, Div150Cred, Div150Multi, Div150Adhoc) all of the modern probabilistic models, from the original ideas of the Probability Ranking Principle [56] to the relatively newer language modelling approaches [57], have as basic building blocks a component that quantifies the importance of a term t in a document d, $F(tf_{t,d})$, and a component that quantifies the specificity of a term, $F'(df_t)$. F and F' are control functions, often $\log$ or rational [58]. The provided data comes with the $tf_{t,d}$ (term frequency) and df_t (document frequency) values for all terms and documents. They are determined per data set basis, and within a data set per image basis, per query basis, and per user basis. To allow reproducibility, we also provide the index files generated with Lucene[4] (can be used directly in Solr or ElasticSearch engines). Some of the best performing systems in the 2014/2015 Retrieving Diverse Social Images Task have used word embeddings to calculated text similarity beyond the tf-df statistics. Therefore, as part of the 2016 data (Div150Adhoc) we have also provided word embeddings for all terms in Wikipedia (*wikiset*). We use the method proposed in [59], i.e., skip-gram with negative-sampling training (SGNS) method in the Word2Vec framework. While this is not the newest method in this category (e.g., the authors in [60] introduced GloVe and reported superior results), independent benchmarking reported [61] show that there is no fundamental performance difference between the recent word embedding models. In fact, based on their experiments, they conclude that the performance gain observed by one model or another is mainly due to the setting of the models' hyper-parameters.

3) User tagging credibility descriptors: (Div150Cred, Div150Multi, Div150Adhoc) our aim is to find a reliable estimation of the overall quality of tag-image content relationships for a user's contributions to Flickr. Using tagging credibility estimates facilitate the design of image retrieval and diversification systems that incorporate the social dimension. It gives an indication about which users are most likely to share relevant images in Flickr. For each Flickr user that has at least one contribution in the *credibilityset*, *devset*, or *testset*, we first retrieve up to 1,000 images and associated metadata (i.e., tags, title, timestamps, etc.). Credibility descriptors are then extracted both from textual and visual data. For Div150Cred, the following descriptors are proposed: *visualScore*: a binary SVM is trained for predicting over 17,000 ImageNet concepts using Overfeat convolutional neural network (CNN) features [62]. The *visualScore* for a user is determined as the average of Flickr tags classification scores, obtained with these models. The higher the prediction score, the higher should be the credibility of user's tags; *faceProportion*: faces are detected using standard tools from OpenCV [63], and *faceProportion* is the percentage of images containing persons for a certain user. Given the scenario of the data, we target images with less faces in foreground, where the target topics are the main focus. The descriptor is related to the relevance of the image; *tagSpecificity*: is the average specificity of a user's tags, where specificity is assessed as the percentage of users that employed those tags for a large corpus of data, namely $\sim$100 million image metadata from 120,000 users; *locationSimilarity*: is the average similarity between the GPS tagged photos of a user and precomputed location models of $1km^2$ cells obtained from the MediaEval 2013 Placing Task [64]. This provides information about the correctness of the geotagging; *photoCount*: is the number of photos shared by a user on Flickr; *uniqueTags*: is the percentage of unique tags in a user's vocabulary; *uploadFrequency*: is the average time between two consecutive uploads on Flickr; *bulkProportion*: is the percentage of identical tags for at least two distinct images. It aims to capture a bulk tagging behaviour.

In the Div150Multi data set, we extended the set of credibility descriptors by including novel estimators, some of which were found to be strong indicators for user tagging credibility in [65]: *meanPhotoViews*: is the average number of image visualizations for a certain user; *meanTitleWordCounts*: is the average number of words used in photo titles for a certain user; *meanTagsPerPhoto*: is the average number of tags used for describing the images of a user; *meanTagRank*: user's tags are sorted by decreasing the amount of their usage (sampled for a large collections of Flickr images). *meanTagRank* is then computed as the average rank of user's tags in this list; *meanImageTagClarity*: is an adaptation of the Image Tag Clarity score described in [66], i.e., the KL-divergence between the tag language model and a collection language model. The adaptation over the initial approach is the use of *tf/idf* language models. This score is an indicator of the various contexts a tag is used in. *meanImageTagClarity* is computed as the average of the clarity scores obtained for a user's tags (for practical reasons, we only consider the appearance in the top 100,000 most frequent tags).

The Div150Adhoc collection proposes most of the previously described credibility descriptors. For this data set, the definition for relevance has slightly changed from previous years, with the introduction of multi-topic queries unrelated to POIs. In particular, *faceProportion* and *locationSimilarity* have been removed as they are no longer valid in the current scenario, while the *visualScore* descriptor was updated to use more effective CNN representations, i.e., we employ now the last fully connected layer of the network in [67].

4) Convolutional neural network descriptors: (Div150Multi, Div150Adhoc) since the initial success, CNN features have been used as universal representations for a variety of image classification and retrieval tasks [68]–[71]. CNNs are commonly used for solving computer vision problems in which train and test concepts are identical.

[4]http://lucene.apache.org/core/

However, if trained with a large number of concepts, generic feature extractors can also be used to characterize other data sets whose concepts overlap the original ones to some extent [68], [72]. Nevertheless, transfer efficiency is reduced whenever the gap between train and test sets is too high and, in such cases, dedicated models should be trained. Therefore, for the Div150Multi and Div150Adhoc collections we provided both general CNN descriptors but also CNN descriptors that were specifically tuned for the POI recognition task. *CNN generic*: a model is trained on 1,000 ImageNet classes. It is provided with the Caffe framework [73]. The descriptor consists of the last fully connected layer of the network (fc7). *CNN adapted*: is inspired by recent domain adaptation work [74]. It is also based on the Caffe framework but uses 1,000 landmark models instead of ImageNet concepts [38]. The descriptor is again the fc7 layer.

C. Ground truth

The presented data sets come with photo relevance and diversity annotations. To disambiguate the diversification need, explicit definitions were provided. They were determined and validated in the community based on the feedback gathered from over 200 respondents during the MediaEval annual community surveys [10]–[13]. *Relevance*: a photo is considered to be relevant if it is *a common photo representation* of the location/of all query concepts at once. Bad quality photos, e.g., severely blurred, out of focus, etc., as well as photos with people as the main subject are not considered relevant. *Diversity*: a set of photos is considered to be diverse if it depicts *different visual characteristics* of the target location/concepts, e.g., sub-locations, temporal information, typical actors/objects, genesis and style information, with a certain *degree of complementarity*, i.e., most of the perceived visual information is different from one photo to another.

Annotations were carried out by trusted assessors (experts) with advanced knowledge of the query characteristics (mainly learned from Internet and Flickr metadata). In particular, to explore differences between expert and non-expert annotations, crowdsourcing annotations were provided for a sample of 50 queries (6,169 photos) from the Div400's *testset*. For the results, the reader is referred to Ionescu et al. [9].

To avoid any bias, annotations were carried out individually on different locations without having the annotators discussing with each other. Following the best practice from the literature [8], [29], we determined the following annotation protocol (annotators used a specially developed software to carry out the process):

Relevance: for each query, the annotators were provided with one photo at time. A reference photo of the query (e.g., a Wikipedia photo) has been displayed during the process, as reference. 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: is annotated only for the photos that were judged as relevant in the previous relevance step. For each query, 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 their visual similarity. The number of clusters was limited to maximum 20 for Div400 and to 25 for Div150Cred, Div150Multi and Div150Adhoc. 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 also provided some keyword tags reflecting their judgments in choosing these particular clusters. The process was also not time restricted.

A summary of the overall annotation statistics is presented in Table II. The relevance ground truth was collected from several annotators leading in the end to 3 distinct annotations per photo. Final relevance ground truth was determined after a lenient majority voting scheme (-1 are disregarded if not in majority). The diversity annotation was also collected from several annotators who annotated distinct parts of the data, leading in the end to 1 annotation per photo. In some cases, a master annotator reviewed once again the annotations.

For measuring the agreement among pairs of annotators, we computed the Kappa statistics. 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 disagreement worse than chance. In general, Kappa values above 0.6 are considered adequate and above 0.8 are considered almost perfect [75]. All the annotations achieve agreement values above 0.6, and maximum 0.85. Also, less than 0.03% of the images were undecided after majority voting which is negligible. On average, more than 67% of the images were considered relevant which is a significant number, allowing to focus on the diversification process. As comparison, crowdsourcing annotations have the agreement significantly lower, i.e., 0.36, which may reflect the variability of the background of the crowd annotators. The number of relevant images is however more or less similar, 69%.

For the diversity annotations, we achieve around 12 clusters per query for Div400, where the maximum number of images per query was 150, around 22 clusters for the tourism scenario where we have up to 300 images per query, and a lower number, i.e., 17 clusters when the queries are more complex (multi-topic). The resulting number of images per cluster is in general in the 8-10 interval. In comparison, crowdsourcing annotations lead to 4.7 clusters per query and 32.5 images per cluster, which shows that crowd workers tend to simplify the process for optimizing the duration of the annotation task.

D. Evaluation Methodology

The classical evaluation metrics from information retrieval are widely used measures for the estimation of search quality [32], [36], [38], [76]. Effective metrics include the Mean Average Precision (MAP) [77], the area under the ROC curve (AUC) [78], and the normalized discounted cumulative gain

TABLE II
DATA GROUND TRUTH STATISTICS (*devset* — DEVELOPMENT DATA, *testset* — TESTING DATA, *credibilityset* — DATA FOR ESTIMATING USER TAGGING CREDIBILITY (RELEVANCE ANNOTATIONS WERE PERFORMED ON A SELECTION OF 50,157 PHOTOS), *single* ($_{st}$) — SINGLE TOPIC QUERIES, *multi* ($_{mt}$) — MULTI-TOPIC QUERIES, *expert* — ANNOTATIONS PERFORMED BY EXPERT ASSESSORS, *crowd* — ANNOTATIONS PERFORMED VIA CROWD SOURCING ON A SELECTION OF 50 QUERIES, 6,169 PHOTOS, *+1* —A MASTER ANNOTATOR REVIEWED ONCE AGAIN THE ANNOTATIONS).

	Div400 (2013)			**Div150Cred (2014)**			**Div150Multi (2015)**				**Div150Adhoc (2016)**		
	devset	*testset*	*testset*	*devset*	*testset*	*credibilityset*	*devset*	*testset*$_{st}$	*testset*$_{mt}$	*credibilityset*	*devset*	*testset*	*credibilityset*
relevance	expert	expert	crowd	expert	expert	expert	expert	expert	expert	expert	expert	expert	expert
#annotators	6	7	175	3	11	9	11	7	5	9	9	9	9
#annotations per image	3	3	3	3	3	3	3	3	3	3	3	3	3
average Kappa	0.64	0.8	0.36	0.85	0.75	0.75	0.77	0.8	0.69	0.75	0.64	0.67	0.75
% relevant images	73.5	65	69	70	67.4	68.6	67.9	63	69	68.6	64.4	50	68.6
% undecided images	0.06	0.04	0.01	0.03	0.01	0.01	0.01	0.01	0	0.01	0.001	0.006	0.01
diversity	expert	expert	crowd	expert	expert	-	expert	expert	expert	-	expert	expert	-
#annotators	3	4	33	2(+1)	3(+1)	-	3(+1)	3(+1)	3(+1)	-	5(+1)	5(+1)	-
#annotations per image	1	1	3	1	1	-	1	1	1	-	1	1	-
avg. #clusters per query	11.6	13.1	4.7	23.17	22.58	-	22.9	20.9	17.2	-	18	16	-
avg. #img. per cluster	6.4	5	32.5	8.89	8.82	-	8.9	9	12.6	-	11	9	-

(NDCG) [79]. However, these metrics do not consider diversity but focus on relevance only. In contrast, several evaluation measures address diversity but do not reflect relevance, such as classical clustering evaluation measures in [29], α-NDCG [80], and user intent aware measures [16]. To reflect both aspects, relevance and diversity, Jang et al. [31] proposed a modified version of the average precision (AP). However, the proposed average diverse precision (ADP) metric defines diversity as a simple dissimilarity measure between ranked images. The missing consideration of the true diversity as provided by a ground truth annotation limits the comparability of the results achieved by different approaches.

The most widely employed evaluation metrics, which account for both relevance and diversity, originate in information retrieval [81], namely: *cluster recall@X* ($CR@X$), *precision@X* ($P@X$), and their harmonic mean $F1@X$ [8], [9], [18], [26], [33], [34], [37]–[39], [41]. $CR@X$ provides the number of clusters from the ground truth that are represented in the top X results and, thus, it reflects the diversification quality of a given image result set. It is defined as: $CR@X = \frac{N}{N_G}$, where N is the number of image clusters represented by the top X ranked images and N_G the total number of image clusters according to the ground truth. Since the clusters in the ground truth consider relevant images only, the relevance of the top X results is implicitly measured by $CR@X$. Nevertheless, $P@X$ provides a more precise view on the relevance of a particular image set. $P@X$ measures the relevance among the top X images and is defined as: $P@X = \frac{N_r}{X}$, where N_r is the number of relevant images in the top X results. We set $X \in \{5, 10, 20, 30, 40, 50\}$.

IV. EXPERIMENTAL RESULTS

The main objective of the performed experiments is to investigate several crucial aspects in the context of image search diversification, such as the capabilities and the evolution of existing systems, the employed features, and the underlying approaches. We analyze the various runs submitted to the *MediaEval 2013–2016 Retrieving Diverse Social Images Tasks* Div400 – 38 runs, Div150Cred – 54 runs, Div150Multi – 59 runs, and Div150Adhoc – 29 runs (180 runs in total). The evaluation of the results is carried out on a common basis as presented in the previous sections.

A. Analysis of the overall performance

In our first experiment we investigate the distribution of the achieved performances for the different data sets. Figure 2 presents a box plot view of the results in terms of precision ($P@X$) and cluster recall ($CR@X$) for the various cutoff values, X=$\{5, 10, 20, 30, 40, 50\}$. Additionally, we report the performance of the initial Flickr retrieval result as a baseline.

The results show that with an increasing number of considered result images, the overall precision only decreases slightly while the cluster recall increases notably. This tendency is independent of the underlying data set and indicates an essential improvement of the diversity of the retrieved results at a comparable relevance level, when more images are considered.

In general, only few submitted runs improved the performance of the initial Flickr ranking in terms of precision ($P@X$), while most of them achieve a comparable or a slightly lower $P@X$ score. This proves the effectiveness of the Flickr retrieval, which would allow to focus more on the diversification part. The increase in diversity of the retrieved image set (depicted by $CR@X$) is notable in comparison to the initial Flick ranking. Additionally, the larger the number of considered images, X, the more evident is the difference in terms of diversification. For example, for the *Div150Adhoc* data set and X=20, the submitted runs achieve an average $CR@20$ score of 0.3997, compared to the Flickr baseline, 0.3609. For X=50, the performance of the submitted runs in terms of $CR@50$ increases to 0.6315 in contrast to the performance of the Flickr baseline, 0.5601.

Some evaluation settings lead to significant spread of the performances of the different runs, indicated by the outliers in Figure 2. For example, while 50% of the submitted runs for

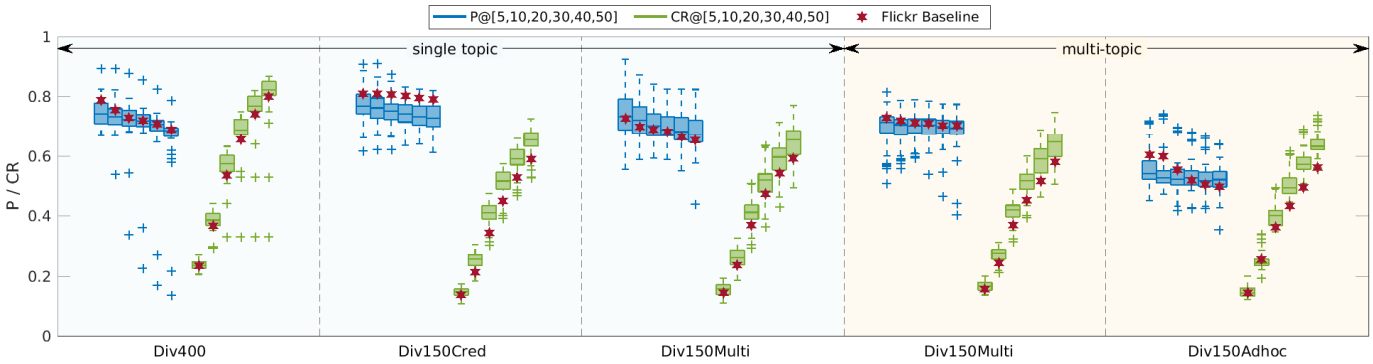


Fig. 2. Performance analysis over the different data sets via boxplot representations of the results (+ are outlier systems, P stands for precision, and CR for cluster recall). Metrics are computed at various cutoff points. Flicks initial retrieval results are provided as baseline.

the *DIV400* data set achieve a $P@50$ score within the narrow range between 0.67 and 0.70, a few runs represent outstanding outliers. For example, Bursuc and Zaharia [82] employ a purely visual-based approach and achieve a $P@50$ score of 0.14 in comparison to the human-based approach by Szűc et al. [83] with the maximum achieved $P@50$ score of 0.79. Nevertheless, since $P@X$ and $CR@X$ are intercorrelated, a proper analysis of potential outliers can only be done in consideration of both evaluation scores. Therefore, we provide more insights into the performance of single approaches in the context of the exploration of the employed features presented in the next section.

Eventually, the different data sets exhibit a varying level of complexity in terms of considered image queries and application scenarios. *Div400*, *Div15Cred*, and a part of the *Div150Multi* cover queries describing single concepts in a tourist scenario. In contrast, the second part of *Div150Multi* and *Div150Adhoc* employ queries representing several concepts in combination, the former data set addressing a tourist scenario and the last one a general-purpose, ad-hoc retrieval system. The differences in the underlying scenarios are captured by the achieved results. A comparison of the performances on the *single topic* data sets and on the *multi-topic* data sets shows a significant difference for both $P@X$ and $CR@X$ for $X=\{10, 20, 30, 40, 50\}$ using the Mann-Whitney-U test [84] ($p \ll 0.001$ for $P@20$ and $p=0.002$ for $CR@20$). The drop in the performance is even more intensified for the general-purpose queries (see *Div150Adhoc*), which is expected, given the higher complexity of the queries.

B. Analysis of the employed features

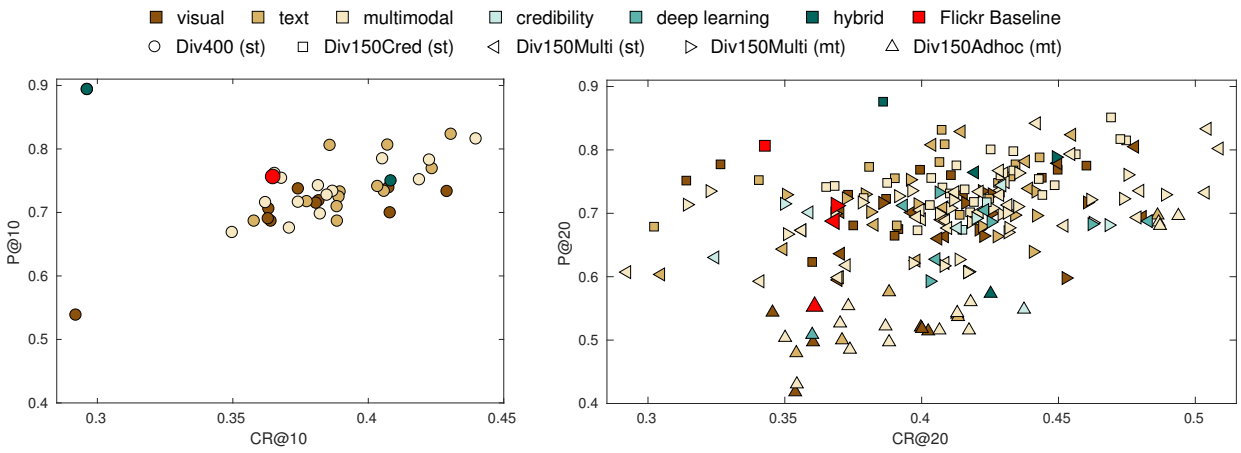
In this experiment we investigate the potential of the various feature types employed (e.g., text, visual, credibility, multimodal, etc.). Figure 3 summarizes the results in terms of the official ranking measures: $CR@10$ and $P@10$ for *Div400* and $CR@20$ and $P@20$ for *Div150Cred*, *Div150Multi*, and *Div15Adhoc*.

The results show that a human-in-the-loop (see *hybrid* or *human-machine*-based approach) commonly leads to a high precision score. For example, Szűcs et al. [83] employ user feedback to organize image retrieval results into clusters of relevant images. The selection of the final image set follows a Round-Robin approach resulting in the highest precision for the *Div400* data set, $P@10=0.89$. Boteanu et al. [85] record user feedback on both relevance and diversity of image result sets to train a Support Vector Machine (SVM) classifier. The approach achieves the highest precision for the *Div15Cred* data set, $P@20=0.88$. Although, user feedback allows for a better assessment of image relevance (and, hence, a higher precision score, in general), the investigated hybrid approaches are notably outperformed by the remaining fully automated techniques in terms of diversification of the final image set, commonly leading to a higher $F1@X$ score as well.

Overall, more than 45% of the the total number of results are reported by *multimodal* approaches, employing combinations of different modalities, such as text, visual, and/or credibility information. Such systems tend to achieve the highest performance in terms of diversification of the final image set independently of the considered data set. Figure 4 shows a detailed view on the performance of the various combinations employed by the investigated systems for the *Div150Cred*, *Div150Multi*, and *Div15Adhoc* data sets. The results show that, currently, the best performing multimodal approaches consider the combination of text and visual information. This combination is widely applied through the different data sets.

More recently, deep learning representations of the textual and/or visual information are commonly considered. While, in general, approaches employing deep learning techniques do not necessarily improve the precision of the retrieved results, they significantly improve the diversity in terms of $CR@20$-score (Mann-Whitney-U test $p=0.004$). For example, Spyromitros-Xioufis et al. [86] achieve the top performance for the *Div150Multi* (st) data set in terms of $F1@20$ score and the highest $CR@20$ score of 0.51 using CNN-based and text-based features for building relevance models and VLAD features to increase the diversity of the retrieved image set. The authors report that the consideration of the deep learning based features improves the overall performance of up to 6%.

Nevertheless, conventional multimodal approaches combining regular text and visual-based features also show remarkable performance. For example, Tollari [87] combines TF-IDF and scalable color in a clustering-based diversification approach. The proposed approach achieves the top performance



(a) Div400 (b) Div150Cred, Div150Multi, Div150Adhoc

Fig. 3. Performance of the employed features (represented with different colors) on the various data sets (represented with different shapes). *st* indicates single topic and *mt* multi-topic data sets.

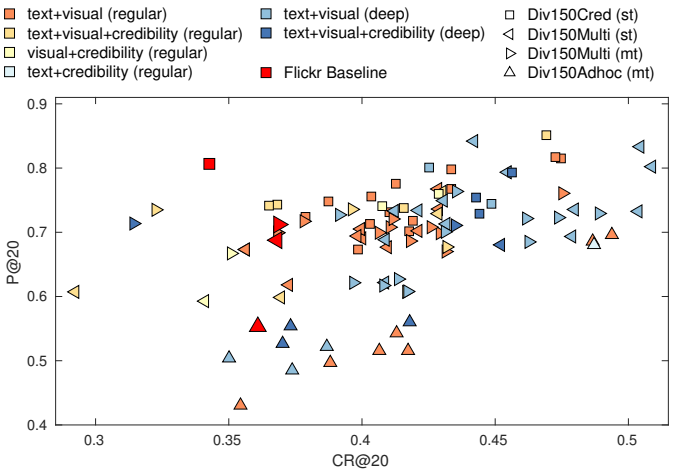


Fig. 4. Detailed view of the performance of the various feature combination considered by the multimodal approaches (represented with different colors) on the various data sets (represented with different shapes). *st* indicates single topic and *mt* multi-topic data sets.

for the *Div15Adhoc* data set improving the Flickr baseline by more than 14% in terms of $F1@20$ score, with $P@20$=0.70 and $CR@20$=0.49 (Flickr: $P@20$=0.55, $CR@20$=0.36).

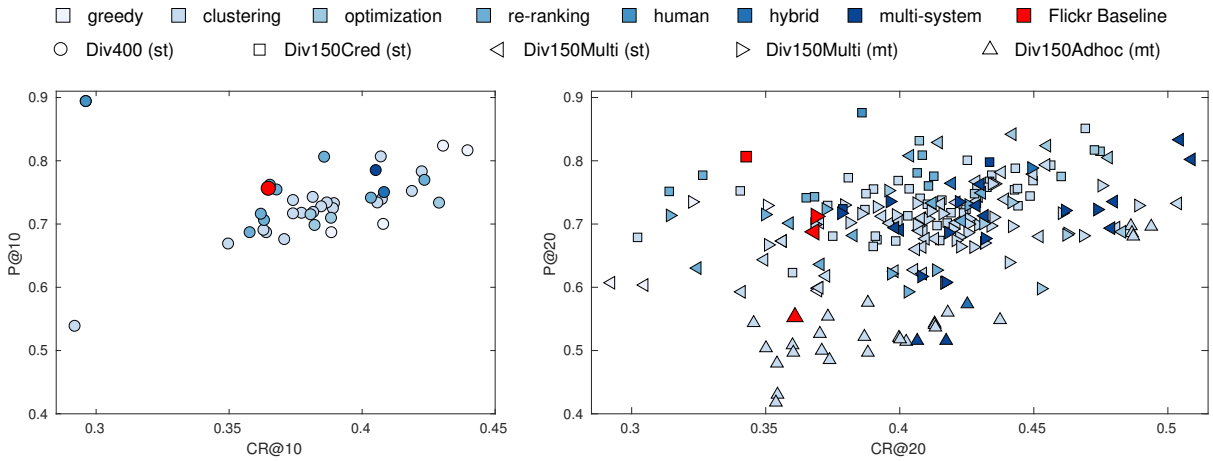
Figure 6 visualizes the underlying fusion techniques employed by approaches combining several features. Please note, that feature combinations do not necessarily consider different modalities (i.e., multimodal approaches) but can also refer to a combination of features of the very same modality. The results show that the minority of the approaches (around 20%) employ late fusion techniques. For example, Ferreira et al. [88], [89] consider a rank aggregation method for rankings generated by different features to improve the original listing of the retrieval results. The approach improves the Flickr baseline for *Div150Multi* (st) and (mt) in terms of diversity ($CR@20$) by more than 5%. Sabetghadam et al. [15] use a weighted linear method and Bayesian inference to combine relevancy and diversity results achieved by text and visual-based features respectively. The approach performs comparable to an early fusion technique, which simply leverages both modalities. $CR@20$ is 0.47 using the late fusion and 0.49 using the early fusion approach on the *Div15Multi* (mt) data set, achieving third and second best results on this data set. The results reflect the overall picture as there is no observable significant difference between the performances of early and late fusion.

C. Analysis of the methods

While in the previous sections we investigated the impact of multiple types of features, we now take a global look at the performance of different approaches to diversity (e.g., greedy, clustering, optimization). Figure 5 gives an overview of the results in terms of the official ranking measures: $CR@10$ and $P@10$ for *Div400* in Figure 5a and $CR@20$ and $P@20$ for *Div150Cred*, *Div150Multi*, and *Div15Adhoc* in Figure 5b.

We first observe that, overall, the clustering based methods are predominant, with 64.2% of the results using a clustering algorithm. We also note that for each particular data set, this approach covers at least 50% of the runs. However, despite being the prevailing technique, it is used by only one top run in terms of precision, for the *Div150Adhoc* data set and two top runs in terms of cluster recall, for the *Div150Adhoc* and *Div150Multi (mt)* data sets. Tollari [87] achieves the best $P@20$ and $CR@20$ scores on *Div150Adhoc* by applying Agglomerative Hierarchical Clustering (AHC) to query results after a re-ranking step. AHC provides a hierarchy of image clusters and the ideal number of clusters varies according to the features used for clustering and the target evaluation measure. On *Div150Multi (mt)*, top $CR@20$ is obtained by Sabetgtham et al. [15], who use ensembles of different configurations of clustering algorithms, features and distance metrics to re-rank a list of results.

Coming on a distant second place in terms of popularity, purely re-ranking based methods are the core of 13.1% of the total number of runs. While this class of approaches does not



(a) Div400 (b) Div150Cred,Div150Multi, Div150Adhoc

Fig. 5. Performance of the employed systems (represented with different colors) on the various data sets (represented with different shapes). *st* indicates single topic and mt multi-topic data sets.

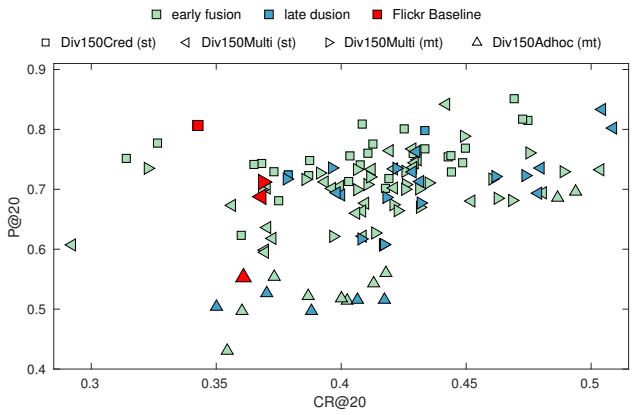


Fig. 6. Detailed view of the performance of the feature fusion approaches (represented with different colors) on the various data sets (represented with different shapes). *st* indicates single topic and mt multi-topic data sets.

provide a top run on any of the data sets, a good result is obtained in [90] on *Div150Multi (st)*. The authors use an iterative algorithm that selects the most different image with respect to all previously selected ones, with the similarity being assessed over visual and textual features. With a CR@20=0.43, a 16.9% relative improvement over the Flickr baseline is observed but it falls behind the best run on this data set [86], which gives CR@20=0.50.

Optimization and multi-system methods account each for 7.6% of the results. By optimizing an utility function defined as a weighted combination of relevance and diversity, the authors in [91] provide the best run for CR@20 on *Div150Cred*. Optimisation techniques are also useful for improving precision, as shown in [92], where the authors obtain the best results in terms of P@20 on *Div150Multi (st)*. The authors formulate the task of diversifying image retrieval results as a subset selection problem and propose to maximize a weighted scoring function composed of submodular functions. The optimization is performed through the use of a sub-gradient descent algorithm [93]. This method also retains good diversity, scoring a $F1$@20=0.56, thus improving the Flickr baseline by 21.3%. Multi-system approaches combine several automated methods. Although not commonly used, they have been found to give consistently good results for diversity. For instance, Spyromitros et al. [86] obtain the highest CR@20 and $F1$@20 on *Div150Multi (st)* using a multimodal ensemble for combining different types of features for relevance detection in a principled manner.

We observe that methods that rely on greedy approaches are less used and account for 4.3% of total runs. They were first used on *Div400* with good results. Jain et al. [94] rank first in terms of CR@10 and $F1$@10 and also attain the highest P@10 among all automated runs. The authors use a Min-Max greedy technique that takes as input a similarity matrix and a pivot image to build the result list. However, greedy methods were not used on the *Div150Cred* and *Div150Adhoc* data sets, while for *Div150Multi (st)* and *(mt)*, the results of these methods are spread among the last 25% runs in terms of CR@20.

Methods that either use human-machine-based or human-in-the-loop (i.e., *hybrid*) approaches are poorly represented. We find only 2 runs for the former and 3 for the latter. As previously observed, these methods lead to a high precision.

From a chronological perspective, we observe a growing trend of multi-system approaches. On the more recently introduced *Div150Multi* and *Div150Adhoc* data sets, they were the second most used methods, after clustering. We also note the time proven consistency of re-ranking methods, which can be found in equal number, with the exception of *Div150Adhoc*, for any test collection. Globally, we observe that each method has outperformed the Flickr baseline at least on one evaluation data set, when looking at the diversity metrics. However, we cannot identify a single method that clearly stands out on any evaluation metric. This finding reinforces the observation that image retrieval diversification can be successfully tackled using approaches from a variety of domains.

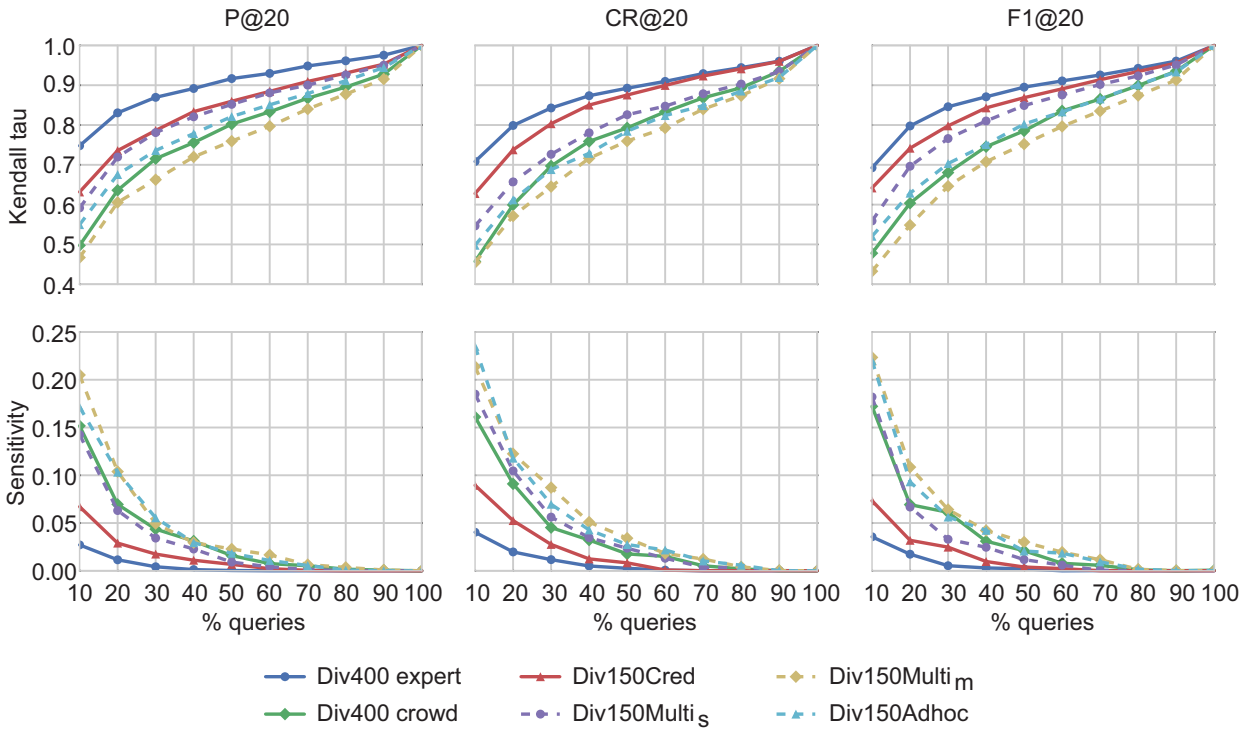


Fig. 7. Kendall correlation tau and sensitivity scores for the proposed evaluation data sets (s - single topic, m - multi-topic).

D. Statistical significance of the results

The final experiment is to analyze the stability of the results. Stability is studied by varying the number of queries $q \in C_Q$ which are used to compute performance, where C_Q is set of queries proposed by collection C. A system is evaluated on a set of queries by averaging the scores of an effectiveness metric (cluster recall in our case). We note by $\lambda_{C_Q,A}$ the average score of an arbitrary effectives metric obtained by system A on C_Q. To compare and rank two systems A and B, the difference between their global performance is taken into account: $\Delta\lambda_{C_Q,AB} = \lambda_{C_Q,A} - \lambda_{C_Q,B}$.

Urbano et al. [95] evaluate several indicators of test collection reliability that have been proposed in the literature. For a collection C_Q, stability evaluation is performed on a pair of subsets from C_Q of the same size, (Q', Q''). Two groups of measures can be distinguished: those who evaluate the rankings and those who take into account the systems' scores. As most of them were found to be relevant evaluators of stability and are correlated among themselves, we chose two of the most commonly employed indicators, one from each family: *Kendall's Rank Correlation* (τ) – this coefficient depends upon the number of inversions of pairs of objects which would be needed to transform the rank induced by $C_{Q'}$ into the rank given by $C_{Q''}$ [96]. It compares the order in which systems are ranked, regardless of the magnitude of the differences $\Delta\lambda_{C_Q,AB}$. It ranges from 1 (perfect correlation) to -1 (inverse correlation). For C to be stable, τ must tend to 1; *Sensitivity* (ρ) – this coefficient is the minimum difference $\Delta\lambda_{C'_Q,AB}/max(\lambda_{C'_Q,A}, \lambda_{C''_Q,B})$ that needs to be observed with Q' such that the differences with Q'' have the same sign at least 95% of the times [95]. C is stable if ρ tends to 0.

Following Voorhees [97], besides evaluating the stability of different test collections, we also investigate the stability of the efficiency metrics that were proposed for evaluating system performance across all collections, i.e., P, CR, $F1$. For each collection presented in this paper and for each evaluation metric, we evaluate stability using random subset pairs $(Q'_{(n)}, Q''_{(n)})$ of various sizes. n represents the percentage of queries retained form C_Q and varies from 10 to 100, with a step of 10. For each n, we randomly sample 100 pairs and report the final results as the average over the 100 trials.

Figure 7 depicts a summary of our stability tests. The results confirm the intuition that the more topics are evaluated, the more stable the rankings are. The values of both stability indicators improve with the number of topics, with a faster pace for $Sensitivity$ compared to Kendall's τ. Urbano et al. [95] conclude that for a collection to be deemed stable it requires to have at least 50 queries. We observe a similar behavior, with our collection presenting stability when using even fewer queries. Among other indicators, it is considered that a collection is stable if $Sensitivity < 0.05$.

From Figure 7 we can see that this criterion is met for all metrics when using at least 40% of the original queries for *Div400* crowd ground truth, *Div150Multi* (single and multi topic), *Div150Adhoc* and at least 20% for *Div400* expert ground truth and *Div150Cred*. This translates to 19, 28, 28, 25, 48 and 24 queries for *Div400* crowd ground truth, *Div150Multi*$_{st}$, *Div150Multi*$_{mt}$, *Div150Adhoc*, *Div400* expert ground truth and *Div150Cred*, respectively. When looking at $Kendall's\ \tau$, a high correlation is commonly considered to appear at a score higher than 0.8. This is equivalent to a stable collection. Except *Div400* expert ground truth, where a high correlation is reached when using only 20% of queries, for

the remaining collections stability is achieved when using at least 40% of queries for *Div150Cred* and at least 60% for the rest. If we compare evaluation metrics in terms of stability, we notice that P@20 is slightly more stable than CR@20, with F1@20, as expecting, being between the first two.

V. Discussion and Conclusions

We introduced a publicly available, common image search diversification benchmark framework that focuses explicitly on social media aspects. It consists of a very rich annotated data, with over 750 single-, multi-topic and ad-hoc queries, 150k images and over 30M image links, metadata, various content descriptors for visual and text modalities. As part of the *MediaEval Retrieving Diverse Social Images Task*, we analyzed four years of results and more than 180 submitted systems, with the objective to provide an in-depth analysis of the crucial aspects of diversification, such as the capabilities and the evolution of existing systems.

A. Lessons learned

Image retrieval capabilities are reliable in terms of precision: analyzing all the submitted systems, only very few managed to improve the precision (of more than 80%) of a state-of-the-art text-based image retrieval system, i.e., Flickr, while most of them achieve a comparable or slightly lower precision. However, default diversification is very low. Therefore, emphasis can be made on the later for improving the usefulness of the image search results.

General purpose diversification capabilities are still a challenge: varying the level of complexity of the considered image queries, from embedding single to multiple concepts in a well defined use case scenario, i.e., tourism, to an ad-hoc, general purpose, use case, shows a significant difference in performance for both precision and diversity. The later is still a challenge to achieve with good performance. However, for the ad-hoc scenario, improvement over the initial retrieval results seems to be more significant, compared to the other data scenarios, which is a promising result.

Human-in-the-loop: it is interesting to notice that approaches harvesting human input, e.g., via crowd sourcing, directly or with hybrid approaches, like pseudo-relevance feedback, tend to improve mostly the relevance of the results, rather than the diversification. The reasoning behind this is the fact that common users who are not familiar with the task would seek for similar images rather than understanding the complex context of the topics, and looking for more meaningful information. Nevertheless, human-in-the-loop is a valuable resource and could be explored to improve the results in the sense that to be better adapted to user needs.

Multimodal approaches are inherently the best performers: almost half of the experimented techniques involve the use of a high diversity of information sources, e.g., visual, social, text. Such systems tend to achieve the highest performance in terms of diversification, independently of the considered data set. Social information plays, in particular, an important role, e.g., via the use of tagging behaviour analysis and credibility. Fusion, i.e., early and late, is traditionally employed to achieve the integration of various modalities, but there is no observable significant difference between their performance, both achieving good results. This means that less computational expensive approaches may be adopted. Deep networks prove again their efficiency, and are able to provide best runners for several data sets. However, it is notable that they are effective as content descriptors rather than classifiers, as they are traditionally used in computer vision applications.

Diversification techniques: a very broad range of approaches has been explored, e.g., 64.2% of the tested approaches use clustering, 13.1% use re-ranking, 7.6% use optimization and multi-system approaches, 4.3% use greedy approaches while very few attempts use human-in-the-loop. From a chronological perspective, one can notice the increase of using multi-system approaches. Overall, there is not a supremacy of one technique over the others, each approach being able to provide best runners at least on one data set. This finding reinforces the observation that diversification can be successfully tackled employing knowledge from a variety of domains and remains an optimization problem between feature design and the decision mechanism.

B. Open questions

Deep learning: has not been actively used in this scenario. So far, almost exclusively, the deep neural networks were employed just as content descriptors for the data. There is the need for network architectures tailored to the diversification task, including the native integration of multi-modal data processing, which are capable of learning the diversification from previous examples. This also means that more annotated data should be made available.

Exploiting user perception of social data: concepts like image visual and social interestingness, memorability, humor, irony, could allow for including the end-user in the diversification process by analyzing the way the selected images are perceived by the actual user. Significant progress has been made in this field, and could be a valuable lead to improve even more the user experience in the diversification task.

Integration with real-world applications: image search diversification is a valid technology which is currently integrated with public search engines. A relevant example is the Google Image Search[5] which integrates a content diversification of the results. Although progress has been significantly made since the first version, research should still be carried out to allow better performance in the ad-hoc scenarios. This involves the possibility of reducing the complexity of the top-end approaches, thus to be adapted to Big Data constraints.

References

[1] T. Kato, "Database architecture for content–based image retrieval," in *Image Storage and Retrieval Systems*, ser. SPIEProc, A. A. Jamberdino and W. Niblack, Eds., vol. 1662, Feb. 1992.

[2] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content–based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, Dec. 2000.

[5]https://www.google.com/imghp

[3] P. G. B. Enser, "Progres in documentation pictorial information retrieval," *Journal of Documentation*, vol. 51, no. 2, 1995.
[4] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.
[5] H. Müller, W. Müller, D. M. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content–based image retrieval: Overview and proposals," *Pattern Recognition Letters*, vol. 22, no. 5, Apr. 2001, special Issue on Image and Video Indexing.
[6] H. Müller, A. Geissbuhler, S. Marchand-Maillet, and P. Clough, "Benchmarking image retrieval applications," in *Proc. of the Conference on Visual Information Systems (VISUAL 2004)*, 2005.
[7] J. R. Smith, "Image retrieval evaluation," in *IEEE Workshop on Content–based Access of Image and Video Libraries (CBAIVL'98)*, 21 1998.
[8] M. L. Paramita, M. Sanderson, and P. Clough, "Diversity in photo retrieval: Overview of the ImageCLEFPhoto task 2009," in *International Conference on Cross-language Evaluation Forum: Multimedia Experiments*, 2010.
[9] B. Ionescu, A. Popescu, A.-L. Radu, and H. Müller, "Result diversification in social image retrieval: a benchmarking framework," *Multimedia Tools and Applications*, vol. 75, no. 2, 2016.
[10] B. Ionescu, A.-L. Radu, M. Menéndez, H. Müller, A. Popescu, and B. Loni, "Div400: A social image retrieval result diversification dataset," in *ACM Multimedia Systems Conference*, 2014.
[11] B. Ionescu, A. L. Gînscă, B. Boteanu, M. Lupu, A. Popescu, and H. Müller, "Div150multi: A social image retrieval result diversification dataset with multi-topic queries," in *International Conference on Multimedia Systems*, 2016.
[12] B. Ionescu, A. Popescu, M. Lupu, A. L. Gînscă, B. Boteanu, and H. Müller, "Div150cred: A social image retrieval result diversification with user tagging credibility dataset," in *ACM Multimedia Systems Conference*, 2015.
[13] B. Ionescu, A. L. Gînscă, M. Zaharieva, B. A. Boteanu, M. Lupu, and H. Müller, "Retrieving diverse social images at MediaEval 2016: Challenge, dataset and evaluation," in *MediaEval 2016 Workshop*, 2016.
[14] T. Tsikrika, A. García Seco de Herrera, and H. Müller, "Assessing the scholarly impact of ImageCLEF," in *CLEF 2011*, ser. Springer Lecture Notes in Computer Science (LNCS), sep 2011.
[15] S. Sabetghadam, J. Palotti, N. Rekabsaz, M. Lupu, and A. Hanbury, "TUW @ MediaEval 2015 Retrieving diverse social images task," in *MediaEval 2015 Workshop*, 2015.
[16] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in *ACM International Conference on Web Search and Data Mining*, 2009.
[17] G. Capannini, F. M. Nardini, R. Perego, and F. Silvestri, "Efficient diversification of web search results," *VLDB Endowment*, vol. 4, no. 7, 2011.
[18] B. Ionescu, A. Popescu, H. Müller, M. Menéndez, and A. L. Radu, "Benchmarking result diversification in social image retrieval," in *IEEE International Conference on Image Processing*, 2014.
[19] K. Zheng, H. Wang, Z. Qi, J. Li, and H. Gao, "A survey of query result diversification," *Knowledge and Information Systems*, 2016.
[20] M. Schedl and D. Hauger, "Tailoring music recommendations to users by considering diversity, mainstreaminess, and novelty," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015.
[21] Y. Shi, X. Zhao, J. Wang, M. Larson, and A. Hanjalic, "Adaptive diversification of recommendation results via latent factor portfolio," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2012.
[22] C. Yu, L. Lakshmanan, and S. Amer-Yahia, "It takes variety to make a world: Diversification in recommender systems," in *International Conference on Extending Database Technology: Advances in Database Technology*, 2009.
[23] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.
[24] Y. Pang, Q. Hao, Y. Yuan, T. Hu, R. Cai, and L. Zhang, "Summarizing tourist destinations by mining user-generated travelogues and photos," *Computer Vision and Image Understanding*, vol. 115, no. 3, 2011.
[25] S. Rudinac, A. Hanjalic, and M. Larson, "Generating visual summaries of geographic areas using community-contributed images," *IEEE Transactions on Multimedia*, vol. 15, no. 4, 2013.
[26] M. Sanderson, J. Tang, T. Arni, and P. Clough, "What else is there? search diversity examined," in *European Conference on Information Retrieval*, 2009.
[27] D.-T. Dang-Nguyen, L. Piras, M. Riegler, M.-T. Tran, L. Zhou, M. Lux, T.-K. Le, V.-T. Ninh, and C. Gurrin, "Overview of ImageCLEFlifelog 2019: Solve my life puzzle and Lifelog Moment Retrieval," in *CLEF2019 Working Notes*, ser. CEUR Workshop Proceedings. Lugano, Switzerland: CEUR-WS.org <http://ceur-ws.org>, September 09-12 2019.
[28] C. L. A. Clarke, N. Craswell, and E. M. Voorhees, "Overview of the TREC 2012 web track," in *Text REtrieval Conference (TREC)*, 2012.
[29] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol, "Visual diversification of image search results," in *International Conference on World Wide Web*, 2009.
[30] M. Wang, K. Yang, X. S. Hua, and H. J. Zhang, "Towards a relevant and diverse search of social images," *IEEE Transactions on Multimedia*, vol. 12, no. 8, 2010.
[31] K. Yang, M. Wang, X.-S. Hua, and H.-J. Zhang, "Tag-based social image search: Toward relevant and diverse results," in *Social Media Modeling and Computing*. Springer, 2011.
[32] Y. Gao, M. Wang, Z. J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Transactions on Image Processing*, vol. 22, no. 1, 2013.
[33] S. Negi and S. Chaudhury, "Identifying diverse set of images in Flickr," in *International Conference on Pattern Recognition*, 2014.
[34] S. Negi, A. Jaju, and S. Chaudhury, "Search result diversification in Flickr," in *International Conference on Communication Systems and Networks (COMSNETS)*, 2016.
[35] L. S. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *International Conference on World Wide Web*, 2008.
[36] A.-L. Radu, B. Ionescu, M. Menéndez, J. Stöttinger, F. Giunchiglia, and A. Angeli, "A hybrid machine-crowd approach to photo retrieval result diversification," in *International Conference on MultiMedia Modeling*, 2014.
[37] D. T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F. G. B. D. Natale, "A hybrid approach for retrieving diverse social images of landmarks," in *IEEE International Conference on Multimedia and Expo*, 2015.
[38] E. Spyromitros-Xioufis, S. Papadopoulos, A. L. Ginsca, A. Popescu, Y. Kompatsiaris, and I. Vlahavas, "Improving diversity in image search via supervised relevance scoring," in *ACM International Conference on Multimedia Retrieval*, 2015.
[39] B. Boteanu, I. Mironică, and B. Ionescu, "Pseudo-relevance feedback diversification of social image retrieval results," *Multimedia Tools and Applications*, 2016.
[40] G. Boato, D.-T. Dang-Nguyen, O. Muratov, N. Alajlan, and F. G. B. De Natale, "Exploiting visual saliency for increasing diversity of image retrieval results," *Multimedia Tools and Applications*, vol. 75, no. 10, 2016.
[41] T. Deselaers, T. Gass, P. Dreuw, and H. Ney, "Jointly optimising relevance and diversity in image retrieval," in *ACM International Conference on Image and Video Retrieval*, 2009.
[42] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The IAPR TC-12 benchmark: A new evaluation resource for visual information systems," in *International Workshop OntoImage*, 2006.
[43] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *International Conference on World Wide Web*, 2011.
[44] A. Olteanu, S. Peshterliev, X. Liu, and K. Aberer, "Web credibility: features exploration and credibility prediction," in *European conference on Advances in Information Retrieval*, 2013.
[45] Y. Yamamoto and K. Tanaka, "ImageAlert: Credibility analysis of text-image pairs on the web," in *ACM Symposium on Applied Computing*, 2011.
[46] F. Benevenuto, T. Rodrigues, A. Veloso, J. Almeida, M. Gonçalves, and V. Almeida, "Practical detection of spammers and content promoters in online video sharing systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 3, 2012.
[47] H. Izadinia, A. Farhadi, A. Hertzmann, and M. D. Hoffman, "Image classification and retrieval from user-supplied tags," *arXiv preprint arXiv:1411.6909*, 2014.
[48] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from national university of singapore," in *ACM International Conference on Image and Video Retrieval*, 2009.
[49] A. L. Gînscă, A. Popescu, B. Ionescu, A. Armagan, and I. Kanellos, "Toward an estimation of user tagging credibility for social image retrieval," in *ACM International Conference on Multimedia*, 2014.
[50] V. de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Transactions on Image Processing*, vol. 18, no. 7, 2009.

[51] O. Ludwig, D. Delgado, V. Goncalves, and U. Nunes, "Trainable classifier-fusion schemes: An application to pedestrian detection," in *International IEEE Conference on Intelligent Transportation Systems*, 2009.
[52] M.Stricker and M. Orengo, "Similarity of color images," *SPIE Storage and Retrieval for Image and Video Databases III*, vol. 2420, 1995.
[53] T. Ojala, M. Pietikäinen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," *IAPR Pattern Recognition*, vol. 1, 1994.
[54] B. Manjunath, J. Ohm, V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, 2001.
[55] X. Tang, "Texture information in run-length matrices," *IEEE Transactions on Image Processing*, vol. 7, no. 11, 1998.
[56] S. E. Robertson, "The probability ranking principle in IR," *Journal of Documentation*, vol. 33, no. 4, 1977.
[57] C. Zhai, "Statistical language models for information retrieval a critical review," *Foundations and Trends in Information Retrieval*, vol. 2, no. 3, 2008.
[58] T. Roelleke, "IR models: foundations and relationships," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2012.
[59] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Annual Conference on Neural Information Processing Systems*, 2013.
[60] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Conference on Empirical Methods in Natural Language Processing*, 2014.
[61] O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Transaction of the Association of Computational Linguists (ACL)*, vol. 3, 2015.
[62] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.
[63] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc., 2008.
[64] A. Popescu and N. Ballas, "CEA LIST's participation at MediaEval 2013 placing task." in *MediaEval 2013 Workshop*, 2013.
[65] A. L. Gînscă, A. Popescu, M. Lupu, A. Iftene, and I. Kanellos, "Evaluating user image tagging credibility," in *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2015.
[66] A. Sun and S. S. Bhowmick, "Image tag clarity: in search of visual-representative tags for social images," in *SIGMM Workshop on Social Media*, 2009.
[67] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.
[68] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014.
[69] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE International Conference on Computer Vision*, 2015.
[70] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *ACM International Conference on Multimedia*, 2014.
[71] J. Yue-Hei Ng, F. Yang, and L. S. Davis, "Exploiting local features from deep networks for image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015.
[72] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
[73] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM International Conference on Multimedia*, 2014.
[74] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *European Conference on Computer Vision*, 2014.
[75] J. Randolph, "Free-marginal multirater kappa (multirater κfree): an alternative to fleiss fixed-marginal multirater kappa," in *Joensuu Learning and Instruction Symposium*, 2005.
[76] L. Yang and A. Hanjalic, "Supervised reranking for web image search," in *ACM International Conference on Multimedia*, 2010.
[77] C. Buckley and E. M. Voorhees, "Retrieval evaluation with incomplete information," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.
[78] B. McFee and G. R. Lanckriet, "Metric learning to rank," in *International Conference on Machine Learning*, 2010.
[79] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, 2002.
[80] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008.
[81] C. X. Zhai, W. W. Cohen, and J. Lafferty, "Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval," in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.
[82] A. Bursuc and T. Zaharia, "ARTEMIS@MediaEval 2013: A content-based image clustering method for public image repositories," in *MediaEval 2013 Workshop*, 2013.
[83] G. Szűcs, Z. Paróczi, and D. M. Vincz, "BMEMTM at MediaEval 2013 Retrieving Diverse Social Images Task: Analysis of Text and Visual Information," in *MediaEval 2013 Workshop*, 2013.
[84] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference (Statistics: a Series of Textbooks and Monogrphs)*, 5th ed. CRC, 2010.
[85] B. Boteanu, I. Mironică, A.-L. Radu, and B. Ionescu, "LAPI@2014 Retrieving Diverse Social Images Task: A relevance feedback diversification perspective," in *MediaEval 2014 Workshop*, 2014.
[86] E. Spyromitros-Xioufis, A. Popescu, S. Papadopoulos, and I. Kompatsiaris, "USEMP: Finding diverse images at MediaEval 2015," in *MediaEval 2015 Workshop*, 2015.
[87] S. Tollari, "UPMC at MediaEval 2016 Retrieving diverse social images task," in *MediaEval 2016 Workshop*, 2016.
[88] R. T. Calumby, Do, V. P. Santana, J. A. V. Munoz, O. A. B. Penatti, L. T. Li, J. Almeida, G. Chiachia, M. A. Gonçalves, and Da, "Recod @ MediaEval 2015: Diverse social images retrieval," in *MediaEval 2015 Workshop*, 2015.
[89] C. D. Ferreira, R. T. Calumby, I. B. A. d. C. Araujo, . C. Dourado, J. A. V. Munoz, O. A. B. Penatti, L. T. Li, J. Almeida, and R. da Silva Torres, "Recod @ MediaEval 2016: Diverse social images retrieval," in *MediaEval 2016 Workshop*, 2016.
[90] A. Lidon, M. Bolaños, M. Seidl, Nieto, P. Radeva, and M. Zeppelzauer, "UPC-UB-STP @ MediaEval 2015 Diversity Task: Iterative Reranking of Relevant Images," in *Working Notes Proc. of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, CEUR-WS.org, ISSN 1613-0073*, 2015.
[91] E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, and I. Vlahavas, "Socialsensor: Finding Diverse Images at MediaEval 2014," in *Working Notes Proc. of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, CEUR-WS.org, ISSN 1613-0073*, 2014.
[92] S. S. Ravindranath, M. Gygli, and L. V. Gool, "ETH-CVL@ MediaEval 2015: Learning Objective Functions for Improved Image Retrieval," in *Working Notes Proc. of the MediaEval 2015 Workshop, Wurzen, Germany, September 14-15, CEUR-WS.org, ISSN 1613-0073*, 2015.
[93] H. Lin and J. A. Bilmes, "Learning mixtures of submodular shells with application to document summarization," *arXiv preprint arXiv:1210.4871*, 2012.
[94] N. Jain, J. Hare, S. Samangooei, J. Preston, J. Davies, D. Dupplaw, and P. Lewis, "Experiments in Diversifying Flickr Result Sets," *Working Notes Proc. of the MediaEval 2013 Workshop CEUR-WS.org, ISSN 1613-0073*, 2013.
[95] J. Urbano, M. Marrero, and D. Martín, "On the measurement of test collection reliability," in *Proc. of SIGIR*. ACM, 2013, pp. 393–402.
[96] H. Abdi, "The kendall rank correlation coefficient," *Encyclopedia of Measurement and Statistics.*, pp. 508–510, 2007.
[97] E. M. Voorhees, "Topic set size redux," in *Proc. of SIGIR*. ACM, 2009, pp. 806–807.