Department of
Information Engineering
and Computer Science DISI

UNIVERSITY OF TRENTO - Italy

DISI - Via Sommarive 5 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

A HYBRID MACHINE-CROWD APPROACH TO PHOTO RETRIEVAL RESULT DIVERSIFICATION

Anca-Livia Radu, Bogdan Ionescu, María Menéndez, Julian Stöttinger, Fausto Giunchiglia, Antonella De Angeli

January 2014

Technical Report # DISI-13-039

A Hybrid Machine-Crowd Approach to Photo Retrieval Result Diversification

Anca-Livia Radu[1,2], Bogdan Ionescu[2], María Menéndez[1], Julian Stöttinger[1], Fausto Giunchiglia[1], and Antonella De Angeli[1]

[1] DISI, University of Trento, 38123 Povo, Italy,
{ancalivia.radu,menendez,julian,fausto,antonella.deangeli}@unitn.it,
[2] LAPI, University "Politehnica" of Bucharest, 061071 Bucharest, Romania,
bionescu@alpha.imag.pub.ro.

Abstract. In this paper we address the issue of optimizing the actual social photo retrieval technology in terms of users' requirements. Typical users are interested in taking possession of accurately relevant-to-the-query and non-redundant images so they can build a correct exhaustive perception over the query. We propose to tackle this issue by combining two approaches previously considered non-overlapping: machine image analysis for a pre-filtering of the initial query results followed by crowd-sourcing for a final refinement. In this mechanism, the machine part plays the role of reducing the time and resource consumption allowing better crowd-sourcing results. The machine technique ensures representativeness in images by performing a re-ranking of all images according to the most common image in the initial noisy set; additionally, diversity is ensured by clustering the images and selecting the best ranked images among the most representative in each cluster. Further, the crowd-sourcing part enforces both representativeness and diversity in images, objectives that are, to a certain extent, out of reach by solely the automated machine technique. The mechanism was validated on more than 25,000 photos retrieved from several common social media platforms, proving the efficiency of this approach.

Keywords: image retrieval, results diversification, crowd-sourcing, image content descriptors, social media.

1 Introduction

The continuously growing number of online personal image collections requires building efficient retrieval systems, e.g., social platforms such as *Panoramio*, *Picasa* or *Flickr* are visited daily by hundreds of millions of users sharing various multimedia resources. Existing social photo search technology is relying mainly on text, image, or more recently on GPS coordinates to provide the user with accurate results for a given query. Retrieval capabilities are however still far from the actual needs of the user. For instance, textual tags tend to be noisy or inaccurate, automatic content descriptors fail to provide high-level understanding of the scene while GPS coordinates capture the position of the photographer and not necessarily the position of the query object.

Social photo retrieval engines focus almost exclusively on the accuracy of the results and may provide the user with near replicas of the query (best matches provide more or less redundant information). However, most of the social media users would expect to retrieve not only representative photos, but also diverse results that can depict the query in a comprehensive and complete manner, covering different aspects of the query. Equally important, focus should be put also on summarizing the query with a small set of images since most of the users commonly browse only the top results.

In this paper we address these particular aspects of photo retrieval and specifically the issue of result diversification. Research on automatic media analysis techniques reached the point where further improvement of retrieval performance requires the use of user expertise. We propose a hybrid machine-human approach that acts as a top layer in the retrieval chain of current social media platforms. Photo search results are firstly filtered using an automated machine analysis step. We designed a new approach that uses a re-ranking scheme for improving representativeness followed by a clustering mechanism that is specifically designed to ensure diversity. Secondly, we employ crowd-sourcing by designing an adapted study with the objective of using human expertise as a final refinement of the results. In this chain, the automated media part has also the role of a pre-filtering step that diminishes the time, pay and cognitive load and implicitly people's work assuring better crowd-sourcing.

The remainder of the paper is organized as follows: Section 2 presents an overview of the literature and Section 3 situates our approach accordingly. Section 4 deals with the automated media analysis while Section 5 deals with the design of the crowd-sourcing studies. Experimental validation is presented in Section 6 and conclusions in Section 7.

2 Previous work

Various approaches have been studied in the context of social media to improve search capabilities and specifically the representativeness and diversification of the results. In the following, we detail some of the most popular approaches that are related to our work, namely: re-ranking, automatic geo-tagging, relevance feedback and crowd-sourcing techniques.

Re-ranking techniques are the closest to our machine analysis part. They attempt to re-order the initial retrieval results by taking advantage of additional information, e.g., the initial query is performed using text while the refinement uses visual information. An example is the approach in [6] that 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 classic Scale-Invariant Feature Transform (SIFT). Another example is the approach in [7]. It defines a retrieved image as representative and diverse based on the following properties: should be representative for a local group in the set, should cover as many distinct groups as possible and should incorporate an arbitrary pre-specified ranking as prior knowledge. To determine these properties, authors propose a unified framework of absorbing Markov chain random walks. A different work [8] defines a criterion to measure the diversity of results in image retrieval and propose three approaches to optimize directly this criterion. The proposed methods have been quantitatively evaluated for 39 queries on 20,000 images from the public ImageCLEF

2008 photo retrieval task which incorporates visual and textual information and also qualitatively on a novel product image data set. In the context of video data, the approach in [9] addresses representativeness and diversity in Internet video retrieval. A video near-duplicate graph representing the visual similarity relationship among videos is built. Then, near-duplicate clusters are identified and ranked based on cluster properties and inter-cluster links. The final results are obtained by selecting a representative video from each ranked cluster.

Re-ranking techniques prove to be an efficient solution to the diversification issue as long as there are enough positive examples among the first returned results of the system. Another limitation is in the fact that a re-ranking system has to learn ”blindly” which examples are relevant based only on the automatic analysis of data contents.

Automatic geo-tagging techniques are not that close related to our approach but provide an interesting alternative to the diversification part. Automatic geo-tagging deals with automatically determining the geographical position of a picture based on its content description, e.g., textual tags, visual descriptors [1]. This allows for identifying similar images, e.g., from similar locations, without employing any GPS information.

Another perspective to our research problem is to take advantage directly of the human expertise. *Relevance Feedback* (RF) techniques attempt to introduce the user in the loop by harvesting feedback about the relevance of the results. This information is used as ground truth for recomputing a better representation of the data needed. Starting from early approaches, such as Rocchio's algorithm that formulates new queries as a weighted sum of the initial results [2], to current machine learning-based techniques (e.g., Support Vector Machines, boosting techniques) that formulate the RF problem as a two-class classification of the negative and positive examples [3]; relevance feedback proved itself efficient in improving the relevance of the results but more limited in improving the diversification. The main limitation of this approach is the need for the end user to be part of the retrieval system, task that cannot be performed automatically.

A much appealing alternative to relevance feedback is to take advantage of the ”crowd” users. *Crowd-sourcing* is defined as “the act of taking a job traditionally performed by a designated agent (e.g., an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call” [4]. Recently, researchers started to explore the potential of crowd-sourcing in tasks which can be subject of individual variations such as annotation of affective videos and perceived similarity between multimedia files [5]. The main benefit of this human-oriented approach consists of humans acting like a computational machine that can be accessed via a computer interface. Although crowd-sourcing shows great potential, issues such as validity, reliability, and quality control are still open to further investigation especially for high complexity tasks.

3 Proposed approach

To ensure representativeness and diversity of photo retrieval results, we propose a hybrid approach that takes advantage of both machine computational power and human expertise in a unified approach that acts as a top layer in the retrieval chain of current social media platforms (this paper continues our preliminary work presented in [10]). It involves the following steps:

1. machine: photos are retrieved using best current retrieval technology, e.g., using text and GPS tags on current social media platforms. These results are numerous (hundreds) and typically contain noisy and redundant information;

2. machine media analysis: automated machine analysis is used to filter the results and reduce the time and resource consumption to allow better crowd-sourcing. We designed a new approach that re-ranks the results according to the similarity to the "most common" image in the set for improving representativeness followed by a clustering mechanism that is specifically designed to ensure the diversification by retaining only the best ranked representative images in each cluster (see Section 4);

3. crowd-sourcing analysis: to bridge further inherent machine semantic gap, crowd-sourcing is used as a final refinement step for selecting high quality diverse and representative images. To cope with the inherent crowd reliability problem, we designed two adapted studies for both representativeness and diversification. The final objective is to summarize the query with very few results which corresponds to the typical user scenario that browses only the top results.

We identify the following contributions to the current state-of-the-art:

− *re-ranking with diversification of the results*: although the issue of results diversification was already studied in the literature, we introduce a new re-ranking scheme that allows for better selection of both relevant and diverse results and provide an in-depth study of the influence of various content descriptors to this task;

− *better understanding of crowd-sourcing capabilities for result diversification*: we provide a crowd-study of the diversification task which has not been yet addressed in the literature and experimentally assesses the reliability of the crowd-sourcing studies for this particular task;

− *hybrid machine-crowd approach*: we study the perspective and the efficiency of including humans in the computational chain by proposing a unified machine-crowd diversification approach where machine plays the role of reducing the time and resource consumption and crowd-sourcing filters high quality diverse and representative images.

The validation of our approach is carried out on a landmark retrieval scenario using a data set of more than 25,000 photos retrieved from several social media platforms.

4 Machine media analysis

We designed a new approach as presented in the sequel as well as in Figure 1 that involves a first step of image ranking in terms of representativeness using the similarity to the rest of the images. Then, all images are clustered and a small number of diverse images coming from different clusters are selected. Finally, a diversity rank is given by means of the dissimilarity to the rest of the selected images. A mediation between the two ranks guarantees the representativeness and diversity in images.

Step 1: for each of the N images in the initial noisy image set, S, we describe the underlying information using visual content descriptors (see Section 6.1). Then, we determine using the features of all images a Synthetic Representative Image Feature (*SRI*) by taking the average of the Euclidean distances between all image features;

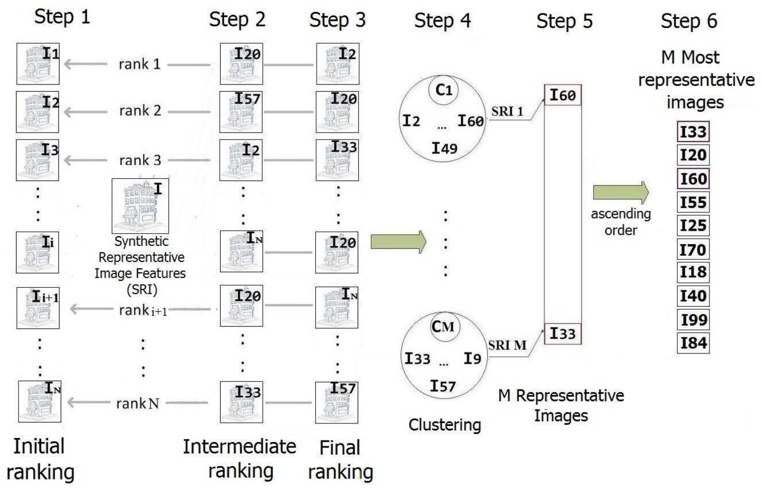


Fig. 1. Proposed re-ranking approach (I represents an image and C denotes the clusters).

Step 2: for each image, I_i, $i = 1, ..., N$, we compute the average of the Euclidean distances to the rest of the images in S which leads to a *N*-dimensional array. The value of *SRI* is subtracted from the array which is sorted in ascending order. The new position of each value will account for the intermediate new rank of its corresponding image;

Step 3: supposing that most representative images are among the first returned, for the final re-ranking we average the two ranks (initial rank and the intermediate one determined in Step 2) which yields *N* average values. Values are again sorted in ascending order and images are re-arranged accordingly;

Step 4: all re-ranked images are clustered in *M* clusters using a k-means approach based on their visual content. Preliminary tests returns good performance for *M* around 30;

Step 5: a SRI_j, $j = 1, ..., M$, is computed as presented in Step 1 and a new re-ranking is performed by re-iterating Steps 1, 2 and 3 over the set of images inside each cluster, C_j. Each cluster's first ranked image is considered to be representative (denoted RI_j);

Step 6: from all RI_j images, we select only a small set of *P* ($P << M$) highest ranked images (ranking according to the final rank computed in Step 3). This will ensure also the diversification of the results.

Experimental validation shows the efficiency of this approach that significantly improves the initial results as well as outperforms another relevant approach from the literature (see Section 6.1).

5 Crowd-sourcing analysis

To bridge further the inherent semantic gap of automatic machine analysis techniques, we use a crowd-sourcing approach. We designed two studies that are adapted to our diversification task, where crowd is involved to refine and improve machine-analysis results with the final goal of determining a high quality set of representative and diverse images. Low monetary cost, reduced annotation time and results close to expert-based approaches [11] are some of the reported advantages of crowd-sourcing that made it very suitable for solving multimedia tasks. However, crowd-sourcing is not a perfect system; not every task can be crowd-sourced and quality control is usually an issue [12].

The reliability of the crowd-sourcing results are analyzed and discussed. The study contributes to the identification of challenges and research directions in crowd-sourcing for image retrieval tasks.

5.1 Crowd-sourcing platform

Several crowd-sourcing platforms appeared in the last years. *Amazon Mechanical Turk* is one of the most popular, although requester accounts are limited to users with a billing address located in United States. An alternative meta crowd-sourcing platform is *Crowdflower*. Jobs created in *Crowdflower* can be published in several crowd-sourcing channels, including *Amazon Mechanical Turk*. Channels usually vary in work force size and contributors' geographical location.

Although judgements can be ordered in different channels, not all functionalities are available via *Crowdflower*. For example, *Amazon Mechanical Turk* allows requesters selecting the workers who can access the job, manually rejecting low quality answers and republishing rejected assignments at no cost. Instead, *Crowdflower* uses an automatic quality control based on gold units. Gold units are unambiguous question for which an answer is provided by the requester. Contributors need to answer at least 4 gold questions with a minimum 70% accuracy to get their answers included in the results. In *Crowdflower* requesters can create jobs, which consist of a data file and units. Units contain the tasks to be performed and are instantiated using the data file. Before ordering, requesters can calibrate the number of judgements per unit, number of judgements per page, and worker pay per page.

5.2 Crowd-sourcing study design

We designed two crowd studies: *Study 1* addresses the representativeness in images provided by the machine-analysis step, while *Study 2* addresses the diversification of the representative images extracted in *Study 1*. The studies were adapted to the monument retrieval scenario used for validating our approach.

Study 1 - representativeness. The representativeness task collected data on the variable locations (i.e., indoor, outdoor), relevance, and representativeness. In this study, relevant pictures contain, partially or entirely, the query monument. Representative images are prototypical outside views of a monument. Relevance is an objective concept indicating the presence/absence of the monument, or part of it, while representativeness is a more subjective concept that might depend on visual context or personal perception.

The task was divided in two parts: first, participants were familiarized with the task and provided with a contextualized visual example of the monument query (*Wikipedia* entry of the monument). Secondly, contributors were asked to answer questions on location, relevance, and representativeness for a given retrieval result.

Study 2 - diversification. The diversification task collected data on perceived diversity among the set of representative pictures of the same monument that were provided by Study 1. Participants assessed visual variation considering the use case scenario of constituting a monument photo album.

As for the previous task, first participants were given an introduction to the task accompanied with some visual examples including a link to the *Wikipedia* page of the monument. Afterwards, contributors were asked to answer whether they would include the provided pictures in the photo album.

6 Experimental results and discussion

To validate our approach we use a data set of more than 25,000 images depicting 94 Italian monument locations, from very famous ones (e.g., "Verona Arena") to lesser known to the grand public (e.g., "Basilica of San Zeno"[3]). Images were retrieved from *Picasa*, *Flickr* and *Panoramio* using both the name of the monument and GPS tags (with a certain radius). For each monument, when available, we retain the first 100 retrieved images per search engine, thus around 300 image in total per monument. To serve as ground truth for validation, each of the images was manually labeled (as being representative or not) by several experts with extensive knowledge of monument locations.

6.1 Validation of machine media analysis

The proposed machine media analysis aims to refine the initial retrieval results by retaining for each query only a small set of around $P = 10$ images (see Step 6 in Section 4) that are both representative and diverse. To assess performance, we use the retrieval precision computed as $tp/(tp + fp)$, where tp is the number of true positives and fp are the false positives.

The first test consisted on determining the influence of the content descriptors on the precision of the results. We experimented with various state-of-the-art approaches:

• *MPEG-7 and related texture and color information*: we compute color autocorrelogram (autocorr.), Color Coherence Vector (CCV), color histogram (hist.), color layout (c.layout), color structure (c.struct), color moments (c.moment), edge histogram (edgehist.), Local Binary Patterns (LBP), Local Ternary Patterns (LTP) and color histogram in [16] (c.n.hist).

• *feature descriptors* that consist of Bag-of-Visual-Words representations of Histogram of oriented Gradients (HoG), Harris corner detector (Harris), STAR features, Maximally Stable Extremal Regions (MSER), Speeded Up Robust Feature (SURF) and Good Features to Track (GOOD). We use 4,000 words dictionaries, experimentally determined after testing different dictionary sizes with various descriptors.

Table 1 summarizes some of the results (we report the global average precision over all the results of the 94 monument queries; descriptors are fused using early fusion). Most interesting is that for this particular set-up very low complexity texture descriptors are able to provide results very close to the use of much more complex feature point

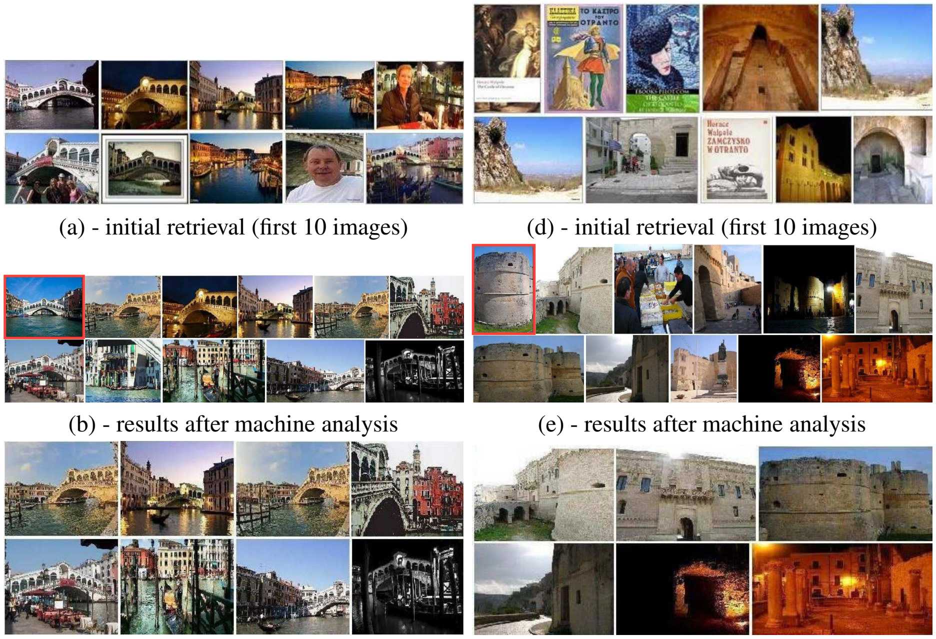
[3] data set is available at `http://www.cubrikproject.eu`, FP7 CUbRIK project.

Table 1. Average precision for various descriptor combinations.

autocorr.	CCV	hist.	c.layout	c.struct	**c.moment**	edgehist.	LBP	LTP
57.70%	58.80%	57.98%	57.70%	60.00%	**60.80%**	56.67%	56.58%	57.68%
Harris	STAR	MSER	SURF	GOOD	c.struct & GOOD	c.struct & HoG	c.moment &GOOD	all color desc.
57.84%	58.87%	59.17%	59.18%	60.39%	60.59%	59.02%	59.71%	58.43%
c.n.hist	HoG	all color desc. & GOOD						
58.33%	58.33%	58.73%						

Table 2. Performance comparison (average precision).

proposed	method in [17]	*Picasa*	*Flickr*	*Panoramio*
60.8%	46.8%	39.47%	53.51%	39.85%


(a) - initial retrieval (first 10 images)

(d) - initial retrieval (first 10 images)

(b) - results after machine analysis

(e) - results after machine analysis

(c) - final results after crowd sourcing

(f) - final results after crowd sourcing

Fig. 2. Output example for the proposed approach: (a)(b)(c) results for query "*Rialto Bridge*" and (d)(e)(f) results for query "*Castle of Otranto*" (image sources: reference *Wikipedia* (image in red rectangle), others *Picasa*, *Flickr* and *Panoramio*).

representations. Therefore, for a higher computational speed, one may go along with a simpler approach without loosing performance. The highest precision is provided by color moments, 60.8%, which are further employed for the machine-crowd integration.

Another interesting result is the limited representative power provided by automatic content descriptors in this context. Regardless the feature used, the disparity of the performance is within a [56%;60%] precision interval. This basically shows the limitation of automated machine analysis and the need for addressing other information sources.

Nevertheless, results achieved by media analysis show significant improvement over the initial retrieval. Table 2 compares these results against the initial retrieval given by the three image search engines and an approach from the literature. The average improvement in precision over initial retrieval is more than 16%. In addition, we achieve an improvement of more then 23% over the approach in [17], that is very promising.

To have a subjective measure of performance, several example results are illustrated in Figure 2 (for visualization reasons, we also display a *Wikipedia* picture of the query - depicted with the red rectangle).

One may observe that compared to the initial images, the automatic machine analysis allows for significant improvement of representativeness and diversity. However, not all the results are perfect, the limited representative power of content descriptors may lead to misclassification, e.g., in Figure 2.(b) some of the initial duplicates are filtered - Figure 2.(a) images 4 and 8, but near-duplicates may appear in the final results, e.g., Figure 2.(b) images 2 and 5.

6.2 Validation of crowd-sourcing analysis

This section reports the results of the representativeness and diversity studies on the images selected by the automated media analysis step. We discuss implications for media analysis and future work. The machine analysis step reduced the initial 25,000 images to only 915 for all the 94 monuments, data set that is further used in the crowd studies.

Crowd-sourcing results on representativeness. The representative study was conducted between 15th and 21st November 2012. Each unit contained one image, which was judged by at least three contributors. Contributors earned 0.07$ per unit. Contributors from 20 different countries located in diverse world regions were allowed to access the job. In total, 228 contributors judged 5,377 units. Due to platform's quality control method, 18% of the units were annotated more than three times. Contributors performed an average of 23.7 units (SD=25.5) with a minimum of 8 and maximum of 174 units. Most of the contributors were located in India (21%), Indonesia (18%), USA (15%), Germany (14%), Canada (9%), Italy (8%), and Morocco (6%).

Reliability analysis was calculated using Kappa statistics that measure the level of agreement among annotators discarding agreement given by chance (values from -1 to 1) [18]. As general guideline, Kappa values higher than 0.6 are considered adequate and above 0.8 are considered almost perfect [11]. There are several variations of Cohen's Kappa depending on categories and raters. For this study, fixed marginal multirater Kappa was used [18]. This variation of Cohen's Kappa is indicated when the number of categories is fixed and there are more than two raters. For analysis purposes, only images annotated exactly three times were considered. In total, 749 images were considered in the reliability analysis. Reliability among annotations achieve a Kappa value of 0.7 for location, 0.44 for relevance and 0.32 for representativeness.

In order to aggregate crowd-sourcing results, the average value among contributors' judgements for location, relevance and representativeness was calculated. First, categorical values (i.e., "Indoor", "Outdoor", "yes", "no") were coded into binary values. "Indoor" judgements for the variable location were coded with 1 and "Outdoor" judgements were coded with 0. For the variables relevance and representativeness, pictures annotated with "yes" were coded with 1 and those annotated with "no" took the value 0. Average values were calculated per image and variable, and mapped onto a binary distribution. For location, pictures with average value equal or above 0.5 were coded as indoor locations. For relevance and representativeness, pictures with average values equal or higher than 0.5 were considered both as related and representative.

Averaged results indicate that 81% of the images depict an outdoor location, 63% contain at least a part of the monument and 57% are representative. As the definition of representativeness used in this paper implies that a picture is representative if it depicts an outdoor scenario, distributions were also calculated considering just images annotated as outdoor. In this case, results show that the percentage of images containing the monument is up to 66%, the percentage of representative pictures is 60%.

Crowd-sourcing results on diversity. Images judged as representative were included in the diversity task, grouped by monument. The diversity study was conducted on November the 26th. In total, 499 images grouped in 82 units (i.e. monuments) were annotated. As the estimated time per unit was similar as in the Study 1, contributors also earned 0.07$ per unit. Units were judged by at least three contributors. In total, 62 contributors participated in 262 units. Number of performed units per contributor varied from 2 to 24 units. Most of the contributors were located in Indonesia (30%), Italy (26%), India (24%), Germany (5%), Brazil (4%), and United Kingdom (4%). Results were averaged and mapped onto a binary distribution using a similar procedure as in Study 1. Aggregated averaged results indicate that: 48% of the monument-grouped images contain all representative and diverse images; some 73% of the grouped images contain at least 75% representative and diverse images, 90% of the group images contain at least 50% representative and diverse images.

Crowd-sourcing is a promising approach for human validation. The results of this study support existing research which identify low costs and reduced annotation time as the advantages of using crowd-sourcing. However crowd-sourcing is not a perfect system, issues such as quality control and reliability of results need further investigation. Automatic quality control methods, as the gold units used in this study, can be a good option since data cleaning is a time-consuming and tedious task [14]. However, they do not always ensure high quality data; since answers can only be rejected at runtime, they may attract more spammers, malicious, and sloppy workers [12] [14].

Reliability of results may vary depending on the level of subjectiveness of the measured variable. For example, annotations for the variable location (i.e. indoor or outdoor) achieve an adequate level of reliability, while the reliability of the annotations for representativeness is quite low. These results suggest that representativeness is a concept subject of individual variations (e.g., visual perception, previous experiences, level of expertise), image features (e.g., perspective, color, and scene composition), or monuments' features (e.g., popularity, distinguishable features, and kind of monument). These issues should be further investigated and considered in the development of methods for aggregation of crowd-sourcing results, since current aggregation methods may underestimate the value of heterogeneous answers [19].

6.3 Validation of the machine-crowd

The final experiment consisted in analyzing the performance of the whole machine-crowd chain. The overall average precision after the crowd step is up to 78% which is an improvement over the machine analysis and initial retrieval results (see Table 2). Several examples are depicted in Figures 2.(c) and 2.(f). In general, if enough representative pictures are provided after the machine analysis, the crowd-sourcing step allows for increasing the diversity among them; while for the case when not enough representative pictures are available, the crowd-sourcing tends to increase the relevance (this is usually due to the fact that these pictures are already highly diverse, but not that relevant).

Overall, the use of machine and human validation allows for better performance than using solely the automated media analysis. The fact that crowd-sourcing acts like a human computational machine allows its integration in the processing chain. Depending on the complexity of the content descriptors (e.g., Bag-of-Visual-Words), crowd-sourcing may yield faster response than the machine analysis, but it is limited in the accuracy of the results for large data sets and therefore running it directly on the initial data is not efficient. Crowd-sourcing is not capable of providing perfect result, despite the use of human expertise. The main reason is that low monetary costs attracts people with limited experience to the task and results may be variable.

7 Conclusions

We addressed the problem of enforcing representativeness and diversity in noisy images sets retrieved from common social image search engines. To this end, we introduced a hybrid approach that combines a pre-filtering step carried out with an automated machine analysis with a crowd-sourcing study for final refinement. The motivation of using the media analysis is also to reduce the workload in the crowd-sourcing tasks for enabling better results. Experimental validation was conducted on more than 25,000 photos in the context of the retrieval of photos with monuments. Results show that automatic media analysis reached the point where further performance improvement requires the use of human intelligence, since regardless the image descriptors use, they are limited to reach only up to 60% precision. Instead, the further use of crowd-sourcing led to an improving of both representativeness and diversity in the final results. Thanks to the initial diversification of the results, after crowd-sourcing some 73% of the grouped images contained a promising number of at least 75% representative and diverse images. Future work will mainly consist on adapting the proposed approach to the large scale media analysis constraints.

Acknowledgment

This research is supported by the CUbRIK project, an IP funded within the FP7/2007-2013 under grant agreement n287704 and by the Romanian Sectoral Operational Programme POSDRU/89/1.5/S/62557.

References

1. A. Rae, P. Kelm, *Working Notes for the Placing Task at MediaEval*, 2012, MediaEval 2012 Workshop, Pisa, Italy, October 4-5, 2012, CEUR-WS.org, ISSN 1613-0073.
2. C. Jordan, C. Watters, *Extending the Rocchio Relevance Feedback Algorithm to Provide Contextual Retrieval*, AWIC04, pages 135-144, 2004.
3. J. L. Elsas, P. Donmez, J. Callan and J. G. Carbonell, *Pairwise Document Classification for Relevance Feedback*, TREC 2009.
4. A.J. Quinn and B. B. Bederson, *Human computation: a survey and taxonomy of a growing field*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11). ACM, New York, NY, USA, 1403-1412, 2011.
5. M. Soleymani and M. Larson. *Crowd-sourcing for affective annotation of video: Development of a viewer-reported boredom corpus*. In SIGIR Workshop on Crowd-sourcing for Search Evaluation, 2010.
6. B. Taneva, M. Kacimi and G. Weikum, *Gathering and ranking photos of named entities with high precision, high recall, and diversity*, ACM on Web search and data mining, pages 431-440, USA, 2010.
7. X. Zhu, A. Goldberg, J. V. Gael and D. Andrzejewski, *Improving Diversity in Ranking using Absorbing Random Walks*, pages 97-104, 2007.
8. T. Deselaers, T. Gass, P. Dreuw and H. Ney, *Jointly optimising relevance and diversity in image retrieval*, ACM on Image and Video Retrieval, pages 39:1–39:8, USA, 2009.
9. 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 Trans. Inf. Syst., vol. 28, pages 22:1-22:27, USA, November 2010.
10. A.-L. Radu, J. Stöttinger, B. Ionescu, M. Menéndez, F. Giunchiglia, "Representativeness and Diversity in Photos via Crowd-Sourced Media Analysis", 10th International Workshop on Adaptive Multimedia Retrieval - AMR 2012, October 24-25, Copenhagen, Denmark, 2012.
11. S. Nowak and S. Rüger, *How reliable are annotations via crowd-sourcing? a study about inter-annotator agreement for multi-label image annotation*, Int. Conf. on Multimedia Information Retrieval, 2010.
12. A. Kittur, E. H. Chi and and B. Suh, *Crowd-sourcing user studies with Mechanical Turk*, SIGCHI Conf. on Human Factors in Computing Systems, pages 453–456, Italy, 2008.
13. D. J. Crandall, L. Backstrom, D. Huttenlocher and J. Kleinberg, *Mapping the world's photos*, Int. Conf. on World Wide Web, pages 761–770, 2009.
14. C. Eickhoff and A. P. de Vries, *Increasing Cheat Robustness of Crowd-sourcing Tasks*, 2012.
15. S. Rudinac, A. Hanjalic and M. Larson, *Finding representative and diverse community contributed images to create visual summaries of geographic areas*, ACM Int. Conf. on Multimedia, pp. 1109-1112, 2011.
16. J. Van De Weijer and C. Schmid, *Applying Color Names to Image Description*, IEEE Int. Conf. on Image Processing, page 493, USA, 2007.
17. L. S. Kennedy and M. Naaman, *Generating diverse and representative image search results for landmarks*, Int. Conf. on World Wide Web, pages 297-306, China, 2008.
18. J. J. Randolph, R. Bednarik and N. Myller, *Author Note: Free-Marginal Multirater Kappa (multirater kfree): An Alternative to Fleiss' Fixed - Marginal Multirater Kappa*.
19. J. A. Noble, *Minority voices of crowd-sourcing: why we should pay attention to every member of the crowd*, ACM Conf. on Computer Supported Cooperative Work Companion, pages 179-182, USA, 2012.
20. J. Li, Q. Ma, Y. Asano and M. Yoshikawa, *Re-ranking by multi-modal relevance feedback for content-based social image retrieval*, APWeb'12, pages 399-410, China, 2012.