

# Pseudo-relevance feedback diversification of social image retrieval results

Bogdan Boteanu<sup>1</sup> · Ionuț Mironică<sup>1</sup> · Bogdan Ionescu<sup>1</sup>

Received: 22 November 2015 / Revised: 8 March 2016 / Accepted: 6 June 2016 /  
Published online: 25 June 2016  
© Springer Science+Business Media New York 2016

**Abstract** In this paper we introduce a novel pseudo-relevance feedback (RF) perspective to social image search results diversification. Traditional RF techniques introduce the user in the processing loop by harvesting feedback about the relevance of the query results. This information is used for recomputing a better representation of the needed data. The novelty of our work is in exploiting the automatic generation of user feedback in a completely unsupervised diversification scenario, where positive and negative examples are used to generate better representations of visual classes in the data. First, user feedback is simulated automatically by selecting positive and negative examples from the initial query results. Then, an unsupervised hierarchical clustering is used to re-group images according to their content. Diversification is finally achieved with a re-ranking approach of the previously achieved clusters. Experimental validation on real-world data from Flickr shows the benefits of this approach achieving very promising results.

**Keywords** Social image search results diversification · Pseudo-relevance feedback · Hierarchical clustering · Image retrieval · MediaEval benchmarking

## 1 Introduction

Nowadays the Internet plays a critical role in our daily life, becoming the main source of information. Besides text, most of these information can also be found in multimedia

---

✉ Bogdan Boteanu  
bboteanu@alpha.imag.pub.ro

Ionuț Mironică  
imironica@imag.pub.ro

Bogdan Ionescu  
bionescu@imag.pub.ro

<sup>1</sup> LAPI, University “Politehnica” of Bucharest, 061071 Bucharest, Romania



**Fig. 1** Search for “Rialto Bridge” on Flickr: images marked by ‘X’ are not relevant to the query, images marked with ‘?’ are ambiguous, images marked by the red rectangle depict redundant results and images in blue rectangles depict examples that are relevant and diverse

files, such as images, videos and audio. As the amount of data increases continuously, it is essential to find our needed information in an effective way. In this context, image retrieval still remains an important topic as a large amount of multimedia data is stored as images on the Internet.

An effective retrieval system should be able to summarize search results and give a global view by surfacing the results which are *relevant* and in the same time they cover different aspects, i.e., they are *diverse*, of the query. Most of the queries involve many possibilities to interpret the interrogations because they may refer to different subtopics, e.g., animals can be part of particular species, cars belong to different types and producers, points of interest can be photographed from different angles and so on. Current state-of-the art retrieval systems, e.g., Flickr<sup>1</sup>, Google Images<sup>2</sup>, focus mainly on relevance and providing real time capabilities for the retrieval process which is often in the detriment of the effectiveness and accuracy of the results. In effect, on many occasions, results tend to be unsatisfactory for the common user. For instance, results can be noisy and not relevant to what the users are searching for; or provide many similar results in terms of visual appearance, e.g., duplicate images. An example in this respect is presented in Fig. 1, namely a Flickr search for “Rialto Bridge”. One can observe that although the search query is more or less common, provided results tend to be redundant (e.g., see images marked by the red rectangle, which are near-duplicates) and many are not relevant to the query or are ambiguous (e.g., see images marked by ‘X’ and ‘?’). A good example of relevant and diverse images are the results marked by the blue rectangles, where we can observe that the images are related to the query and are also visually different, i.e., they depict different angles, different moments of the day, etc.

By increasing the relevance of the results, one can expect to provide more accurate information to the user; whereas by widening also the pool of possible results

<sup>1</sup><https://www.flickr.com/>

<sup>2</sup><https://www.google.com/imghp>

(i.e., diversify) one can increase even more the effectiveness of the system, e.g., by responding to the needs of different users, by tackling queries with unclear information needs, by widening the pool of possible results, reducing the number/redundancy of the returned items, and so on. In general, relevance and diversity tend to be antinomic, i.e., increasing solely the relevance will tend to provide many near duplicates while too much diversification may result in losing relevant items [18]. The key of the entire diversification process is to mitigate these two components. Although relevance was studied more thoroughly in existing literature than diversification [11, 34, 42] and even though a considerable amount of diversification literature exists (in particular in the area of text retrieval, from where the concept emerged), the topic is currently still an open research issue, especially for social media [35, 38, 50, 52].

In this paper we investigate a particular class of methods, i.e., relevance feedback (RF), which were developed initially for increasing relevance only, but show promising results also for diversification. RF take advantage directly of the human expertise in the retrieval process [23, 57, 60]. A general RF scenario consists of the following steps. For a certain retrieval query, the user gives his opinion by marking the results as relevant or irrelevant. Then, the system automatically computes a better representation of the information needed based on this feedback and the retrieval is further refined. This process can go on several cycles. Although effective in many scenarios, there are some drawbacks of this approach. If to consider a real-world retrieval scenario it is not realistic to assume that relevance judgments are available at the time of the retrieval and the process of completing the feedback can be a time consuming task for the user. In the same time, traditional relevance feedback harvests only relevance information, which limits its capabilities to improving only these aspects. A particular case of RF is the pseudo-relevance feedback [3, 24]. Pseudo-RF represents an automatic method to simulate the users's feedback, so that the results are improved without any interaction with the system. The main idea behind this concept is to assume that a small number of top-ranked items in the initial retrieval results are relevant, and use them to re-rank the results. Many experiments such as the results from Trecvid [32] and ImageClef [16] show improvement in retrieval system's performances using this concept.

In this work, we introduce a new pseudo-relevance feedback approach in an unsupervised diversification scenario based on a hierarchical clustering algorithm and a re-ranking scheme. The automatic generation of user feedback, i.e., the positive and negative examples, is used to generate better representations of visual classes in the data. To disambiguate the diversification need, we focus in this paper on a specific use case, i.e., the search for images with tourist landmarks. Landmark locations are popular with social media platforms and also benefit from spatio-temporal and visual invariance which makes them suitable for benchmarking scenarios.

The remainder of the paper is organized as following. Section 2 presents the overview of the literature on diversification and relevance feedback and positions our contribution. The proposed approach is explained in Section 3. Sections 4 and 5 deal with the experimental setup and validation, while Section 6 concludes the paper.

## 2 Previous work

The concept of search result diversification emerged naturally from the text community, which is responsible for the very first publicly available indexing systems; therefore most of the work on diversification is concentrated around text-based search [12, 53, 63]. However,

with the success of the Internet and social media, the problem has been extended to media information and many diversification approaches have been introduced, e.g., use of visual descriptors [50, 52, 54], user tagging credibility information [13], video features [4].

From their early stages until present, many directions have been investigated. A class of approaches are the *election-based* techniques, where images vote for other images to capture similarity and to achieve diversification. For instance, Leuken et al. [52] propose a method where every image decides by which image (besides itself) it is best represented, accumulating a score which in the end determines chances of the image of being elected as representative. The remaining images are assigned to the representative images using some similarity rules, e.g., a Nearest Neighbor approach.

Another direction is to use *clustering techniques*, in which diversification is achieved by employing some clustering/classification algorithms; some of the commonly used being Hierarchical Clustering, k-means, Support Vector Machines (SVM), etc. For instance, Dang-Nguyen et al. [13] propose for diversification a method based on Hierarchical Clustering algorithm, followed by isolated cluster removing and re-ranking. They employ face, blur, GPS and user credibility-based pre-filtering to remove irrelevant images. Data is represented by text (TF-IDF), visual and user credibility descriptors. Another example is the approach of Sarac and Duygulu [43] which uses for relevance a binary SVM classifier and for diversification k-means clustering. The output is formed by the centroids of each class and by images which are closest to the centroids, according to a k-Nearest Neighbors approach. Data is represented by visual descriptors.

Some other approaches use *re-ranking techniques*, which attempt to compute a new ranking of the results using a different modality than the one used in the initial retrieval, which is typically text. This is often achieved using the visual similarities between retrieved images and other representative images. For instance, Pedronette et al. [36] propose a re-ranking approach based on the k-Nearest Neighbors of a query image, which define the contextual information for the query image. Data is represented by visual and textual descriptors. Spyromitros-Xioufis et al. [44] propose a re-ranking strategy based on the scores of a utility function which jointly considers relevance and diversity. They use a supervised classification model that capture a query and application-specific notion of relevance. Ksibi et al. [26] use an 'adaptive diverse relevance ranking' which automatically predicts an effective trade-off between relevance scores and diversity scores according to the query ambiguity level. Relevance score is assessed using a random walk with restart process and diversity score is a combination of semantic and visual scores.

Apart from the "automated" techniques, more and more attention is paid to the concept of "human in the loop", where human input/feedback is included in the processing chain. In a first attempt, this was carried out by conducting user studies on the systems' results. As this approach is far from being able to perform in real time, other perspectives had to be considered. One lead is to take advantage of the potential of *crowdsourcing* platforms [45] in which humans (i.e., users around the world) act like a computational machine that can be accessed via a computer interface. Issues like validity and reliability are present here and opened to further investigation, since the involvement of untrained people in high complexity tasks may not always bring good results. Another perspective is the use of *relevance feedback*, in which a part of the information is assessed by the user and further provided to the system. This concept has proven to increase retrieval accuracy and gives more personalized results for the user. The relevance feedback problem can be formulated either as a two class classification of the negative and positive samples; or as an one class classification problem, i.e., separate positive samples by negative samples. Using a graphic interface, a human assessor usually annotates a part of the retrieved

images as relevant or irrelevant, according to a relevance criteria. The number of the images to be annotated is much lower than the number of the results. These images are further processed so that the result is a re-ranked list of relevant results. Basically, the algorithm learns from the provided data and tries to refine the whole list of results in order to obtain better precision in terms of relevance. A part of these results can be further evaluated again by the user in a second iteration. The process can be repeated until there are no big variations in terms of precision or until a minimum desired degree of performance is achieved.

One of the earliest and most successful RF algorithms is the Rocchio's algorithm [39] (which is still used at the present time). Using the set of relevant and irrelevant documents selected from the current user relevance feedback window, the Rocchio's algorithm modifies the features of the initial query by adding the features of positive examples and subtracting the features of negative examples to the original feature. Another relevant approach is the Relevance Feature Estimation (RFE) algorithm [40]. It assumes that for a given query, according to the user's subjective judgment, some specific features may be more important than others. A re-weighting strategy is adopted which analyzes the relevant objects in order to understand which dimensions are more important than others in determining "what makes a result relevant". Features with higher variance with respect to the relevant queries lead to lower importance factors than elements with reduced variation. Machine learning techniques can be used also to this task. These approaches involve a training step using the data collected from the user. Then all the results are ranked according to the classifiers's confidence level [28], or classified as relevant or irrelevant depending on some output functions [61]. Some of the most successful techniques use Support Vector Machines [28], Nearest Neighbor approaches [52], classification trees, e.g., use of Random Forests [1]; or boosting techniques, e.g., AdaBoost [61].

Although they provide good results, the main disadvantage of RF is in the need of human expertise, which may not be available at any time and is also time consuming, due to the manual annotation process. A good tradeoff between performance and time are the pseudo-relevance feedback techniques, where the user is substituted by automatic methods of selecting the positive and negative examples. Some of the first approaches in this respect were text-based [5], where first top-ranked documents are the positive examples and the last documents are the negative ones. Positive examples and negative examples can be used as a training set for machine-learning techniques [14, 29] and for multimedia retrieval [62]. This perspective seems very promising especially in the context of the current state of the art retrieval systems, which are able to provide an initial high relevance of the results.

In this paper, we exploit a novel perspective of the diversification via the use of pseudo-RF and clustering/re-ranking. Traditional RF introduces the user in the loop by harvesting feedback about the relevance of the search results. This information is used as ground truth for recomputing a better representation of the data needed. We propose an alternative pseudo-relevance solution for rendering this process completely automatic, while still maintaining a high relevance. First, user feedback is simulated automatically by selecting positive and negative examples from the initial query results. Then, a completely automatic unsupervised hierarchical clustering is used to re-group images according to their content. Diversification is finally achieved with a re-ranking approach of the previously achieved clusters. Experimentation on real-world Flickr data shows the benefits of this approach which despite its complete automation is able to outperform or provide comparable results to user-based relevance feedback approaches as well as other consecrated diversification schemes.

### 3 Proposed approach

The proposed approach operates on top of an existing retrieval system and works as a re-ranking step that refines the initial query results. It involves the following steps:

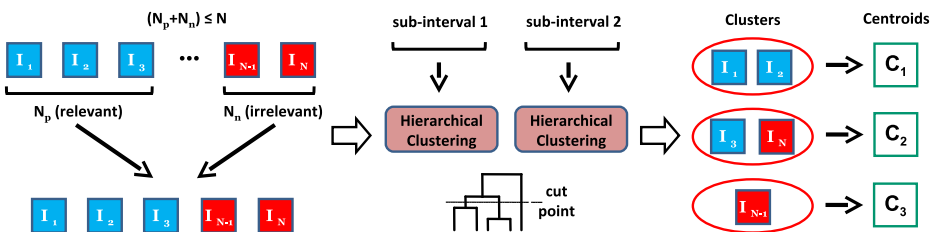
1. *examples selection*: positive and negative examples are selected from the query results. The total number of the selected examples is split in equal sub-intervals (see Section 3.1);
2. *clustering*: for each sub-interval, an unsupervised classification step is used to cluster these examples (see Section 3.2);
3. *merging*: because different sub-intervals may contain similar clusters, clusters are further merged using a folding approach. For this step, we adapted the approach in [52] (see Section 3.3);
4. *pruning*: each resulting cluster is further evaluated based on the number of the relevant and irrelevant images within. This step will ensure the relevance of the refined images (see Section 3.4);
5. *diversification*: following the initial ranking of the retrieved images, cluster images are progressively selected to form the refined diversified query results (see Section 3.5).

Each of the processing steps is detailed in the following sections.

#### 3.1 Selection of positive and negative examples

The first step of the proposed approach consists of selecting a number of positive and negative query examples. Instead of using a classic relevance feedback strategy where the user is supposed to provide these examples, we use a pseudo-relevance feedback approach [3]. In general, current retrieval technology is capable of providing high quality results in terms of relevance. Therefore, we can assume that most of the very first returned results, tend to be relevant for the query. In contrast, the very last of the results are highly likely to be noisy and irrelevant. For instance, in the case of Flickr’s state-of-the-art retrieval system, results in [19, 20] show that, in average, among the first 50 returned images, at least 37 images are relevant to the query, i.e., 75.37 % — estimate obtained for 549 location related queries where the query was formulated as keywords using the location’s name.

Inline with this, we retain the first  $N_p$  images from the initial  $N$ , ranked images as positive examples and the last  $N_n$  images as negative examples (we denote this as *experimental assumption 1*, i.e., the use of first images and last images as positive and negative examples, respectively, to improve the overall system’s performance — see Fig. 2). This leads to a total number of  $N_t$  examples (i.e.,  $N_p + N_n$ ) that constitutes the ground truth. To address



**Fig. 2** Example selection and clustering: selection of positive and negative examples ( $N_p$  and  $N_n$ ),  $N$  is the total number of returned images; followed by sub-intervals clustering and centroid representation



the borderline case when the number of returned images is lower than  $N_t$ , we adopt the following approximations where the positive examples are a fraction of the number of initial ranking images:

$$N'_p = \frac{N}{N_t} \cdot N_p, \quad N'_n = N - N'_p \quad (1)$$

The immediate advantage of this strategy is in the complete automation of the relevance feedback process. No real user interaction is actually required, which reduces significantly the processing time as well as the need for conducting complex user studies. Moreover, this step has impact over the diversification part of the approach, because improving the relevance will also improve diversification.

### 3.2 Clustering

Equipped with the ground truth, we use a clustering strategy to group images with similar content. We select a Hierarchical Clustering (HC) [25] scheme that proved highly effective in various diversification scenarios [6, 19]. The HC scheme uses the "bottom up" approach (i.e., agglomerative). The clustering algorithm works as follows. First, each of the images is assigned to an individual cluster. Then, pairs of new formed clusters are merged according to a similarity measure (i.e., distance metric). The process is repeated until it ends with a single, overall, cluster. Besides its low complexity, HC has the advantage of providing a dendrogram of classes by grouping images iteratively based on a certain distance metric. This allows for adapting the number of output classes to the target scenario based on the selection of a cutting point in the dendrogram.

We apply HC only to the selected positive and negative examples, according to the following strategy. All other unselected images are ignored in the following steps. First, we apply a non-overlapping sliding window approach by dividing the selected examples into small sub-intervals of size  $s$  (see Fig. 2). Preliminary experiments [2] showed that HC separates the data more precisely if we have a small number of images as input, e.g., tens rather than having hundreds. The clustering of less images is better in terms of visual information, i.e., the obtained clusters tend to contain only similar images and images from one cluster are different to images from another cluster (we denote this as *experimental assumption 2*, i.e., *the use of sub-interval division improves the overall system's performance*).

Second, as shown in Fig. 2, the actual HC is applied independently to each of the sub-intervals and the cutoff points of the resulting dendrograms are chosen based on the inconsistency coefficient threshold  $N_c$ .

### 3.3 Merging

The next step is to merge similar image clusters. Because of the sub-division applied, some resulting clusters may be similar, e.g., images of one cluster formed in the sub-interval  $W_i$  may be visually similar to other images in another cluster from sub-interval  $W_j$  ( $i \neq j$ ). The key is to merge these clusters into a single one, representing the sub-interval  $W_{ij}$ .

To do so, we adapted the folding approach in [52] which diversifies a list of retrieved images by first constructing a set  $S$  of representative images and then by assigning the rest of the images to images of set  $S$ . The construction of the representative images is done dynamically, starting from the first retrieved image which is by default assigned to set  $S$ . Next retrieved image is considered representative if it is sufficiently dissimilar to all the images in this set, i.e., the distance is greater than the mean of distances to the mean image of

entire set of all images. In this case, the image is assigned to set  $S$  and the process is repeated until there are no representative images left. Each of the remaining images is assigned to a representative image using the Nearest Neighbor approach. This last step represents the construction of clusters around the representative images, so that similar images are grouped in the same cluster.

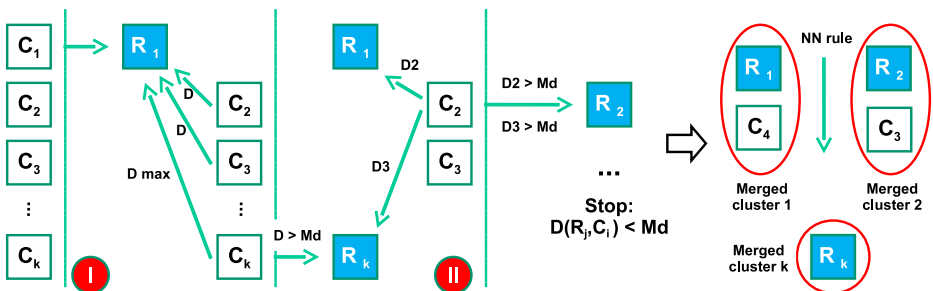
We apply the same principle for our merging problem. Instead of using a ranked list of images, we use the centroids of each cluster (computed by averaging all images) obtained after HC (see Fig. 2). After running the folding approach in [52] we end up with similar centroids grouped together, while the (sufficiently) different centroids will remain in different clusters. Then, all the images corresponding to each centroid are assigned to their new cluster (see Fig. 3). Although cluster similarity highly depends on the size of the clusters, this assumption will hold for this case of sufficiently small clusters that compose a small sub-interval. The reason we choose this approach is the power of discrimination of the method. Despite its high complexity, the method is able to separate well the classes [52] and thus to produce very good results. Therefore, we use Hierarchical Clustering to cluster a large number of images and then a more complex method with high power of discrimination applied on a reduced number of images. This leads to a good combination in terms of performance and resources consumed.

### 3.4 Pruning

In the following, we adopt a supplementary pruning step. A cluster is declared irrelevant if it contains only negative examples or if the number of negative examples is higher than of the positive ones, namely:

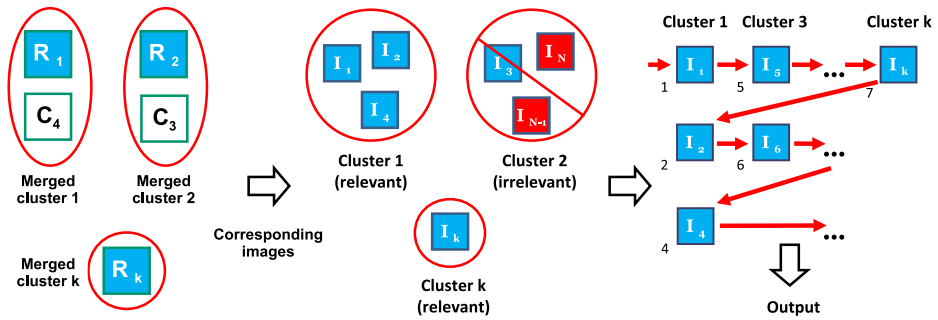
$$N_n^{(i)} \geq \frac{N^{(i)}}{2} \tag{2}$$

where  $N_n^{(i)}$  is the number of negative examples in cluster  $i$  and  $N^{(i)}$  is the total number of examples in cluster  $i$  (we denote this as *experimental assumption 3*, i.e., the adoption of un-relevant clusters improves the overall system’s performance — see Fig. 4). This assumption is based on the fact that cluster images are supposed to be similar with each other. Therefore, if more than half of the present examples are negative, there is a high probability that all the images are in fact negative examples and were assigned wrongly to the positive category.



**Fig. 3** Merging similar image clusters:  $C_i$  is the centroid of cluster  $i$ ,  $R_j$  is the centroid selected as representative,  $D$  represents the Euclidean distance between the centroids of the clusters and  $Md$  represents the average distance computed between all images and the average image. Centroids are merged according to their similarity





**Fig. 4** Cluster pruning and diversification

Experimental results show that this is a good approximation (see Section 5.2). The advantage of this step consists in removing images with a high probability of being un-relevant, and consequently in improving relevance and diversification.

### 3.5 Diversification

The final step is the actual diversification of the results, which is a round-robin approach. To improve also the relevance, we take into account the initial ranking of the results, as the first retrieved images have a higher probability to be relevant than the last ones. To enforce the diversity, we restrict the output to contain at least one image from each HC generated cluster. The algorithm is the following.

First, for each of the HC output relevant clusters (the clusters declared as irrelevant are discarded from diversification), the images are sorted according to their initial ranking, so that the first image in a cluster is the one which has the highest rank in the initial retrieval results. Then, considering the order described above and starting with the first cluster, i.e., the class labeled as the first one by the HC scheme, we select as output each cluster first ranked image. This leads to  $k$  images, where  $k$  is the total number of clusters. The process is repeated iteratively, and clusters are covered again by selecting the second ranked images, third ranked and so on. If in a certain cluster, the number of images is lower than the number of the current iteration (e.g., in the third iteration we attempt to select images from a cluster that has only two images), then that particular cluster is disregarded in the current and further iterations. Finally, the process is repeated until the desired number of images is achieved (see Fig. 4).

## 4 Experimental Setup

### 4.1 Datasets

To test our approach, we selected two publicly available social image retrieval diversification datasets, namely Div150Multi [21], and Div150Cred [19]:

- *Div150Multi dataset*: contains 292 location related queries (e.g., museums, bridges, parks, monuments) with up to 300 photos per query and associated metadata retrieved from Flickr using Flickr’s default “relevance” algorithm (a total of 86,769 images). Images are annotated for both relevance and diversity by human assessors. In particular,

for diversity, images are clustered into similar appearance classes. This dataset proposes a development set containing 153 of the queries (45,575 images) and an evaluation set of 139 queries (41,394 images). The evaluation data contains results obtained with two kind of queries: 'single-topic' — only the name of a location is used; and 'multi-topic' — combined queries related to events and states associated with locations, e.g., "Oktoberfest in Munich", "Bucharest in winter";

- *Div150Cred dataset*: similar to Div150Multi, it contains 153 location related queries with up to 300 photos per query and associated metadata retrieved from Flickr (a total of 45,375 images). Images are also annotated for both relevance and diversity by human assessors and the data is divided into a development set containing 30 of the queries (8,923 images) and an evaluation set of 123 queries (36,452 images).

To be able to compare to the benchmarking results reported on these datasets, for evaluation we used the same experimenting conditions, i.e., the method parameters were optimized on the development data and the actual evaluation is performed on the evaluation set.

## 4.2 Evaluation metrics

To assess performance, we compute the standard cluster recall at a cutoff at  $X$  images ( $CR@X$ ) [35] and the precision at  $X$  images ( $P@X$ ), given by:

$$CR@X = \frac{N}{N_{gt}}, \quad P@X = \frac{N_r}{X} \quad (3)$$

where  $N$  is the number of image clusters represented in the first  $X$  ranked images and  $N_{gt}$  is the total number of image clusters from the ground truth ( $N_{gt}$  is limited to a maximum of 25 clusters for the Div150Multi [21], and Div150Cred [19] datasets),  $N_r$  is the number of relevant images among the first  $X$  ranked results.  $CR@X$  assesses how many clusters from the ground truth are represented among the top  $X$  results provided by the retrieval system. Since clusters are made up of relevant photos only, relevance of the top  $X$  results is implicitly measured by  $CR@X$ , along with diversity. To have a clearer view of relevance,  $P@X$  measures the number of relevant photos among the top  $X$  results. To account for an overall assessment of both diversity and precision, we also report  $F1@X$ , i.e., the harmonic mean of  $CR@X$  and  $P@X$ . Results are reported as overall average values over all the queries in the dataset.

## 4.3 Image content representation

In the clustering process images are represented by content descriptors. Although the approach is not dependent on a certain type of description scheme, the choice of the descriptors influence significantly the results and should be adapted to the specificity of the evaluation data.

Given the specificity of the task, i.e., diversifying social media contents, we tested a broad category of visual and text-based descriptors, which are known to perform well in image retrieval tasks:

- *global color naming histogram* (CN, 11 values): maps colors to 11 universal color names: "black", "blue", "brown", "grey", "green", "orange", "pink", "purple", "red", "white", and "yellow" [15];
- *global Histogram of Oriented Gradients* (HoG, 81 values): represents the HoG feature computed on 3 by 3 image regions [30];

- *global color moments* computed on the HSV Color Space (CM, 9 values): represent the first three central moments of an image color distribution: mean, standard deviation and skewness [46];
- *global Locally Binary Patterns* (LBP, 16 values) computed on gray scale representation of the image [33];
- *global Color Structure Descriptor* (CSD, 64 values): represents the MPEG-7 Color Structure Descriptor computed on the HMMD color space [31];
- *global statistics on gray level Run Length Matrix* (GLRLM, 44 dimensions): represent 11 statistics computed on gray level run-length matrices for 4 directions: Short Run Emphasis, Long Run Emphasis, Gray-Level Non-uniformity, Run Length Non-uniformity, Run Percentage, Low Gray-Level Run Emphasis, High Gray-Level Run Emphasis, Short Run Low Gray-Level Emphasis, Short Run High Gray-Level Emphasis, Long Run Low Gray-Level Emphasis, Long Run High Gray-Level Emphasis [51];
- global descriptor which is obtained by the concatenation of all values and *spatial pyramid representations* of these (denoted with 3x3): each of the previous descriptors is computed also locally; the image is divided into 3 by 3 non-overlapping blocks and descriptors are computed on each patch;
- *generic convolutional neural network* based descriptors (CNN gen, 4,096 values): based on the reference CNN model provided along with the Caffe framework, model which is learned with the 1,000 ImageNet classes used during the ImageNet challenge. The descriptors are extracted from the last fully connected layer of the network (named fc7) [44];
- *adapted convolutional neural network* based descriptor (CNN ad, 4,096 values): based on a CNN model obtained with an identical architecture to that of the Caffe reference model, which is learned with 1,000 tourist points of interest classes whose images were automatically collected from the Web. Similar to CNN generic, the descriptors are extracted from the last fully connected layer of the network (named fc7) [44];
- histogram representations of *term frequency* (TF), *document frequency* (DF) and *term frequency - inverse document frequency* (TF-IDF) information computed on image metadata (descriptor average size of 500 values) [58];
- *user image annotation credibility* information (Credibility, 9 values): an estimation of the global quality of tag-image content relationships for a user's contributions [21, 37, 47].

Descriptors were experimented individually or in combination. Fusion is carried out with an early fusion approach preceded by a max-min normalization.

#### 4.4 Pre-filtering

To improve more the relevance of the results, we pass the initial retrieved images through several pre-filtering steps, namely:

- first, we use the Viola-Jones [55] *face detector* to filter out images with persons as the main subject. These images are in general irrelevant for the common user. The output of the filter for an image consists in a number of pairs of coordinates, indicating where it is most likely to have the face of a person. If this number is greater than a threshold  $T_f$ , then the image is considered to contain faces and it is removed;
- second, we use an *image blur detector* to remove the out of focus images. Regardless their content, severely blurred images are in general not satisfactory results for a query.

We use the aggregation of 10 state-of-the-art blur indicators<sup>3</sup>, namely: Brenner's indicator, graylevel variance, normalized GLV, energy of gradient, thresholded gradient, energy of Laplacian, modified Laplacian, variance of Laplacian, Tenengrad, and sum of wavelet coefficients. An image is rejected if the average of the normalized values is lower than a threshold  $T_b$ ;

- finally, in particular for these datasets, we use a *GPS-based filter*. The filter rejects the images that are positioned too far away from the query location, and therefore which cannot be relevant shots for that location. We use a tolerance radius of  $T_d$  Kms. For accurate results, distance between GPS coordinates is computed using the Harvesine formula [56]. Images for which the GPS information is not available are passed by the filter.

**Limitations** The proposed filters are more or less domain restricted, in particular adapted to the tourism search scenario. However, these filters are a pre-processing step which can be turned on or off if the method is to be applied to other domains or data and not specifically a part of the algorithms.

## 5 Experimental results

This section presents the experimental validation results on the two datasets. We have conducted the following experiments: Section 5.1 deals with method's parameter tuning; Section 5.2 deals with the experimentation of the adopted experimental assumptions of the proposed approach; Section 5.3 compares the proposed method to reference relevance feedback approaches from the literature; Section 5.4 situates our results in the context of the state-of-the-art diversification approaches; Section 5.5 concludes the experimentation by discussing several visual diversification examples.

### 5.1 Parameter tuning

The performance of the proposed approach depends on the choice of several parameters. We run the following experiments on the Div150Multi dataset.

First test consists of tuning the pre-filters to leverage the relevance. We employed filters only on *single – topic* set, as on *multi – topic* set pre-filtering doesn't improve performance in terms of relevance. For instance, face-based filter is not employed because of the specificity the dataset, for which the majority of the photos represent people attending location-related events (e.g., carnivals, concerts) or representing crowds or groups of demonstrators etc. These photos are considered an exception, thus they are relevant, because people attending such events are natural part of the events themselves. We tested the impact of this type of pre-filtering over the performance in terms of relevance and we noticed that, instead of increasing or preserving it, it is even worse, by significantly reducing the precision. For the blur detector, we didn't noticed any performance improvement, thus we

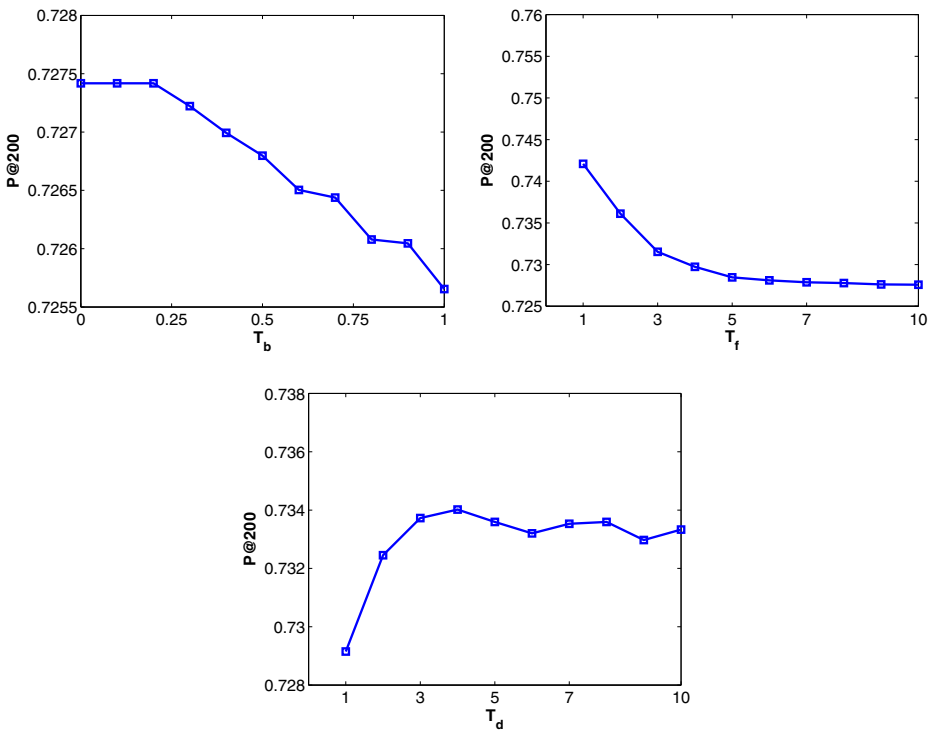
---

<sup>3</sup>An implementation of the used blur indicators is available at <http://www.mathworks.com/matlabcentral/fileexchange/27314-focus-measure/content/fmeasure/fmeasure.m>

concluded that there were no blurred photos in the dataset, and regard to the distance-based filter, the dataset was not provided with GPS coordinates, as there were not available for the majority of the queries (e.g., “rainbow over the lake”, “animals at zoo”, “snow in the city”).

We vary  $T_b$  from 0 to 0.6 with a step of 0.1,  $T_f$  from 1 to 3 with a step of 1 and  $T_d$  from 0 to 5 with a step of 1 (see Section 4.4) so that we obtain maximum precision on the development data. Precision was calculated on the first 200 images retrieved and best performance is achieved by taking  $T_b = 0$ ,  $T_f = 1$ , and  $T_d = 3$ , which leads to  $P@200 = 0.7472$ . Without pre-filtering, the precision is 0.7273, which means that we obtain an improvement of around 2 percentage points.

We also conducted some experiments to see the influence of the parameters for each filter separately, starting from the baseline case when the results are retrieved by Flickr. Thus, we varied  $T_b$  from 0 to 1 with a step of 0.1,  $T_f$  from 1 to 10 with a step of 1, and  $T_d$  from 1 to 10, with a step of 1. Results are presented in Fig. 5. The most important observation is that the biggest performance improvement is given by the face-based filter, i.e., around 2 percentage points (no filtering means  $T_f = Inf$ , because the filter rejects the images with at least  $T_f$  faces), followed by distance-based filter - 0.5 percentage points (no filtering means  $T_d = Inf$ , because the filter rejects the images with distance equal or greater than  $T_d$ ) and blur detector - 0.2 percentage points (no filtering means  $T_b = 1$ ). We can notice that the overall tendency for each parameter setting is to decrease the performance in terms of



**Fig. 5** Influence of the filters’ parameters:  $T_b$ ,  $T_f$ ,  $T_d$

precision as the values increase, because they get closer to the baseline case where no filter is employed.

The following test consists of determining the best descriptor - HC parameter combination (i.e., choice of  $N_p$ ,  $N_n$ ,  $N_c$  — see Section 3.1). In what concerns the descriptors used, we tested individual descriptors and most of the early fusion combinations of the implemented descriptors. For brevity reasons, we restrict the presentation to the use of only one descriptor for each modality and combinations of all of them for each type (e.g., all visual, all textual), as the other combinations didn't lead to any more significant results. We experiment with varying  $N_p$  (number of positive examples) from 100 to 280 with a step of 50,  $N_n$  (number of negative examples) from 0 to 20 with a step of 10, and  $N_c$  (which is related to the number of diversity classes) from 0.7 to 1.3 with a step of 0.2. We use the HC's standard Euclidean distance metric. We experiment also with various sizes of the image sub-interval,  $s$ , namely 10, 20 and 40 images, and we select the value for the best results, which is  $s = 20$  images. We use this value in the next experiments. Table 1 presents the results for the optimal configurations (pre-filtering is used only on single-topic data). We report the metrics at a cutoff at 20 images which is the official metric for this dataset. Overall, the best results in terms of  $F1@20$  are obtained with CNN descriptors and  $N_p = 100$ ,  $N_n = 20$ ,  $N_c = 0.7$ , which yielded  $F1@20 = 0.5444$ . CNN descriptors proved to be very efficient for diversification, maintaining in the same time a good performance for relevance. We use these configurations for the next experiments.

The final experiment was conducted for selecting the best distance metric - cluster merging approach for the HC. For the previously selected parameters, we experiment now with various distance metrics, namely: Euclidean, sEuclidean, cityblock, Minkowski, Chebyshev, cosine, correlation, Spearman, Hamming, and Jaccard [25]; combined with several modalities of computing distances between clusters, namely: single [48] - shortest distance of all distances between each image of one cluster and images of another; complete [27] - furthest distance of all distances; average - unweighted average distance, the resulting distance is the average of all distances between images of clusters; centroid - centroid distance which is the distance between the centroids of two clusters; median - the resulting distance is the median distance of all distances between images of two clusters; Ward [59] - inner

**Table 1** Diversification results for various descriptor - HC parameter combinations (best results are represented in bold)

Set	Modality	Descriptor	$N_p$ - $N_n$ - $N_c$	$P@20$	$CR@20$	$F@20$
Single-topic	Visual	CNN ad	100-20-0.7	<b>0.7065</b>	<b>0.465</b>	<b>0.5435</b>
	Text	Text all	200-20-1.3	0.6819	0.4217	0.5074
	Credibility	Credibility	100-0-1.1	0.7007	0.4404	0.5247
Multi-topic	Visual	CNN all	200-10-0.7	<b>0.7079</b>	<b>0.4666</b>	<b>0.5472</b>
	Text	Text all	250-20-1.3	0.685	0.4365	0.5165
	Credibility	Credibility	280-20-0.9	0.685	0.4313	0.5123
Overall	Visual	CNN ad	100-10-0.7	<b>0.7004</b>	<b>0.4689</b>	<b>0.5444</b>
	Text	TF	200-20-1.3	0.6845	0.4209	0.5057
	Credibility	Credibility	150-20-1.3	0.6838	0.4297	0.5122

squared distance based on minimum variance algorithm; and weighted - weighted average distance, the resulting distance is a combination of all distances between images of each cluster. In all cases, the minimum resulting distance criteria is used to merge two clusters and form a larger one. The best combination results are reported in Table 2. Overall, the highest  $F1@20$  is achieved for Correlation-Ward combination,  $F1@20 = 0.5471$ .

We also experiment the influence of the parameters  $N_p$ ,  $N_n$  and  $N_c$  over the performance of the proposed approach. Thus, we start from the best performance configuration presented in Table 1 and vary separately each parameter while keeping constant the other ones, i.e., we vary  $N_p$  from 10 to 280 with a step of 10,  $N_n$  from 0 to 20 with a step of 5 and  $N_c$  from 0 to 1.3 with a step of 0.1. Results are reported for  $F1$  metric score and presented for each set in Fig. 6. As one can see, selecting the number of classes has the biggest influence over the overall performance, i.e., a variation of around 6 percentage points for *Multi – topic* set and around 3 percentage points for *Single – topic* set. The performance is approximately constant for  $N_c$  from 0.1 to 0.7, then it decreases abruptly for greater values. The second most important parameter is the number of positive examples followed by the number of negative examples. Another interesting observation is related to the number of positive examples. While for the *Multi – topic* set a greater number means better performance, on the *Single – topic* set a greater number of positive examples affects negatively the overall performance. This could be an indicator that within the *Single – topic* set only the first 100 images are in general truly relevant, according to the ground truth, while for *Multi – topic* this number is higher.

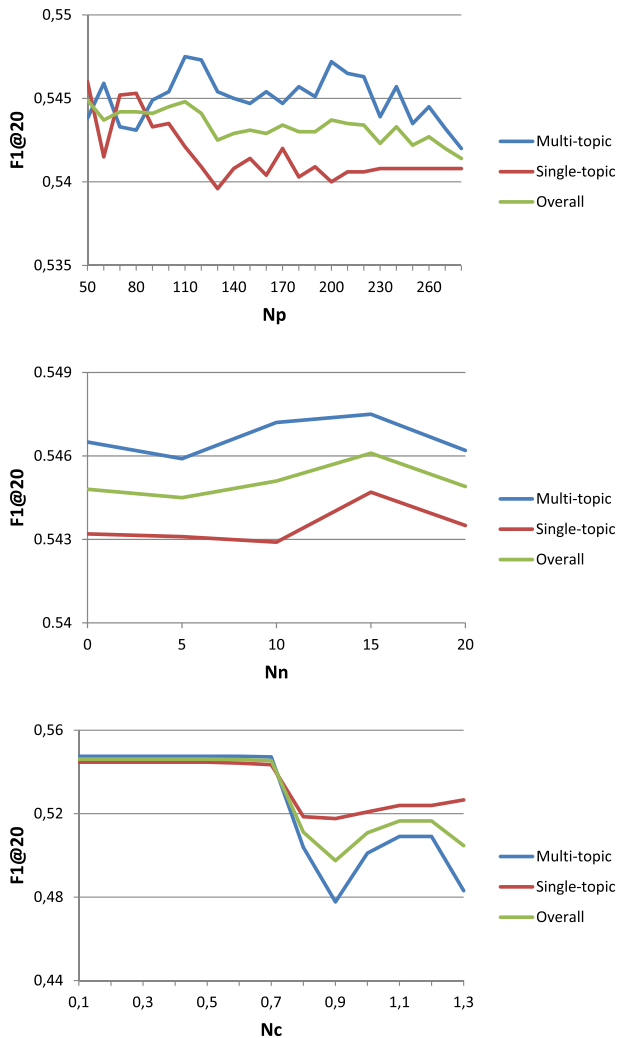
## 5.2 Experimental assumptions

In this section we conduct several experiments in order to assess the advantages of some of the main assumptions of the proposed approach. Experiments are performed on the Div150Multi dataset which is the biggest of the two and using previous best performance parameter tuning. Of course, a proof of these experimental assumptions would require a rigorous significance test of the results. However, this preliminary validation shows promising results.

**Table 2** Diversification results for various HC metrics - merging approaches combinations (best results are represented in bold)

Set	Distance-merging approach	P@20	CR@20	F1@20
Single-topic	Euclidean-single (default)	0.7065	0.465	0.5435
	Euclidean-median	0.7036	0.459	0.538
	Cosine-centroid	<b>0.713</b>	<b>0.4714</b>	<b>0.5509</b>
Multi-topic	Euclidean-single (default)	0.7079	0.4666	0.5472
	Euclidean-median	0.7029	0.4528	0.5348
	Cosine-single	<b>0.7171</b>	<b>0.4731</b>	<b>0.5533</b>
Overall	Euclidean-single (default)	<b>0.7072</b>	0.4658	0.5454
	Euclidean-median	0.7032	0.4558	0.5364
	Correlation-ward	0.7043	<b>0.4708</b>	<b>0.5471</b>





**Fig. 6** Influence of the number of the proposed method's parameters:  $N_p$ ,  $N_n$ ,  $N_c$

**Experimental assumption 1** the adoption of the positive and negative examples (see Section 3.1). We test the impact of taking as negative examples the very last of the returned images. Therefore, for the best results, we set  $N_n = 0$  (no negative examples). In this case we achieve  $F1@20 = 0.5432$  on *single – topic* data and  $F1@20 = 0.5465$  on *multi – topic* data. This is lower than the use of negative examples by around 0.7 percentage points in both cases. Another test is to assess the number of un-relevant images in the set of negative examples. This can be an indicator of how important is the number of negative examples. For 20 negative examples, 19.37 of them are in average un-relevant in the case of *Single – topic* set, while for *Multi – topic* set 9.95 out of 10 images are un-relevant, which means a ratio of more than 95 percent in both cases.

**Experimental assumption 2** the adoption of interval subdivision (see Section 3.2). We test the usefulness of dividing the set of the retrieved images into small sub-intervals and the aggregation of similar clusters. We took into consideration two scenarios: (i) the data is split and HC is applied for each sub-interval, then we merge the similar clusters; (ii) HC is applied on all retrieved images. In the second case, we obtained  $F1@20 = 0.5582$  on *single – topic* data and  $F1@20 = 0.5381$  on *multi – topic* data. The adoption of the interval division leads to a slightly lower performance for *single – topic* data (e.g., around 0.5 percentage points lower  $F1@20$ ) and a significant improvement for the *multi – topic* data, i.e., an increase of 1.5 percentage points; which overall is better than the scenario without this step.

**Experimental assumption 3** the adoption of irrelevant classes in HC (see Section 3.4). To test the usefulness of building the irrelevant classes and thus removing them from the results, we experiment by considering all the classes as relevant. This yields an  $F1@20 = 0.541$  on *single – topic* data and  $F1@20 = 0.546$  on *multi – topic* data, which is almost 1, respectively 0.7 percentage points lower than the result with removing irrelevant classes.

**Diversification improvement** we test the benefits in terms of solely diversification of the proposed method. To do so, we consider the ideal case when the selected positive examples are relevant (selected from the relevance ground truth). This way, we enforce to see the diversification capabilities. First, we consider as baseline the case when all images are relevant and they are retrieved according to their initial rank from Flickr and no diversification is employed. The results in this case are presented in Table 3. Then, we evaluate the performance of the proposed method using the relevance ground truth, i.e., the positive examples are restricted to be genuine relevant. The idea behind this experiment is to compare the improvement in terms of diversification given by the proposed method when the relevance remains constant. In this case, we select all images which are relevant in their initial order and feed them to the proposed method. For the image selection we set  $N_p = 300$  and  $N_n = 0$ , i.e., all images are positive examples, and for HC we varied  $N_c$  from 0.7 to 1.3 with a step of 0.2. In terms of modality, we employed all types of descriptors: visual, textual

**Table 3** Performance of the proposed method in the ideal case when the positive examples are selected from the relevance ground truth (descriptors were combine using early fusion approach)

Set	Modality	Descriptor	$N_p-N_n-N_c$	$P@20$	$CR@20$	$F@20$
Single-topic	Baseline	-	-	0.9884	0.5018	0.6
	Visual	CNN all	300-0-0.9	0.9884	<b>0.5951</b>	<b>0.7294</b>
	Text	Text all	300-0-1.3	0.9884	0.5289	0.6744
	Credibility	Credibility	300-0-1.3	0.9884	0.5615	0.7016
Multi-topic	Baseline	-	-	0.9843	0.4893	0.6294
	Visual	CNN ad	300-0-0.9	0.9843	<b>0.6219</b>	<b>0.7478</b>
	Text	Text all	300-0-0.7	0.9843	0.5694	0.7028
	Credibility	Credibility	300-0-0.7	0.9843	0.5862	0.7138
Overall	Baseline	-	-	0.9863	0.4955	0.6396
	Visual	CNN ad	300-0-0.9	0.9863	<b>0.6031</b>	<b>0.7341</b>
	Text	Text all	300-0-0.7	0.9863	0.5495	0.6885
	Credibility	Credibility	300-0-0.9	0.9863	0.5687	0.706

and credibility (see Section 4.3), or combinations of them, using early fusion. The results are presented in Table 3. By analyzing the results we can observe that the proposed method improves diversification on all sets, regardless of the modality used. Overall, in terms of  $F1@20$ , we achieved the lowest value of 0.6885 using text descriptors, which is still almost 4.9 percentage points higher compared to baseline, and the highest value of 0.7341 using CNN descriptors, which is almost 9.4 percentage points higher compared to baseline. This means that the proposed method achieves better diversification over the baseline, when the input is the same.

### 5.3 Comparison to other relevance feedback approaches

In this experiment we compare our results against other relevance feedback approaches from the literature, namely: Rocchio [39], Relevance Feature Estimation [40] (RFE) and some classification-based approaches: Support Vector Machines (SVM) [28], AdaBoost [61], Random Forests [1] and k-Nearest Neighbors [17]. User relevance feedback is simulated with the images' ground truth in a window of 10 images (which is a typical browsing scenario). We experimented with two situations: (i) feedback is simulated with the relevance ground truth (*relevance*); (ii) feedback is simulated with the diversity ground truth by selecting diverse images in the initial feedback window (*diversity*) — this should allow for more emphasis on the diversification. All the approaches were tuned to best performing parameters (parameters - descriptors). Descriptors were combined using early fusion approach.

First, we compare to the 2015 MediaEval Retrieving Diverse Social Images benchmarking [21] (Div150Multi dataset). For the proposed approach we use the combination of parameters which yielded best performance, i.e., descriptor- $N_p$ - $N_n$ - $N_c$ -distance-merging approach. These are: for *single - topic* CNN ad-100-20-0.7-cosine-centroid, for *multi - topic* CNN all-200-10-0.7-cosine-single, and *overall* CNN ad-100-10-0.7-correlation-ward. Results are presented in Table 4. The first observation is the fact that the best results are obtained for all methods using CNN descriptors or a combination of them, except for Random Forests. In terms of F1 metric score, best result ( $F1@20 = 0.5976$ ) is achieved on *multi - topic* set, using Support Vector Machines (SVM) [28] with linear kernel, using diversity ground truth. The same method performs best on *overall* set ( $F1@20 = 0.5688$ ) using the same type of Ground Truth. On *single - topic* set, best performance is achieved using Rocchio's algorithm with diversity ground truth ( $F1@20 = 0.5656$ ). Another observation is that, overall, best performance is achieved using the diversity ground truth,  $F1@20 = 0.5976$ . An explanation for this can be that in the first retrieved images there are many relevant and diverse images, and the SVM-RF algorithm only retrieve on the first position the same images, because they have the greatest confidence score.

In terms of relevance, the proposed method doesn't perform best on any of the sets, and this can be a consequence to the fact that there are images which are not relevant in the first retrieved images and the only major improvement is the use of the three filters. However, in terms of diversification, the proposed method performs best on *single - topic* and *overall* sets, with  $CR@20 = 0.4714$  and  $CR@20 = 0.4708$ , which is an improvement of almost 1 and 0.3, respectively, percentage points over the best RF method, i.e., Rocchio ( $CR@20 = 0.4618$  and  $CR@20 = 0.4681$  respectively). On *multi - topic* set, the proposed method is on the third place, with  $CR@20 = 0.4731$ , after SVM and Rocchio approaches. The good diversification is achieved due to the fact that the main steps of the proposed method consist in diversification strategies: interval division, Hierarchical Clustering, merging, and diversification.

**Table 4** Comparison to relevance feedback approaches on Div150Multi dataset [21] (best results are represented in bold)

Set	RF Approach	Feedback	Descriptor	$P@20$	$CR@20$	$F@20$
Single-topic	Proposed	pseudo-rel.	CNN ad	0.713	<b>0.4714</b>	0.5509
	Rocchio [39]	relevance	Text all-CNN all	0.7906	0.4467	0.5563
	Rocchio [39]	diversity	Text all-CNN all	0.7768	0.4618	<b>0.5656</b>
	RFE [40]	relevance	Vis all-CNN all	0.7572	0.4262	0.5335
	RFE [40]	diversity	Text all-CNN all	0.7029	0.4569	0.5419
	SVM [28]	relevance	Text all-CNN all	<b>0.821</b>	0.4432	0.5603
	SVM [28]	diversity	Text all-CNN all	0.7594	0.4533	0.5513
	Random Forests [1]	relevance	LBP3x3	0.8036	0.393	0.506
	Random Forests [1]	diversity	LBP3x3	0.7413	0.4315	0.5287
	AdaBoost [61]	relevance	CNN ad	0.6812	0.4352	0.5128
	AdaBoost [61]	diversity	CNN ad	0.7022	0.4292	0.511
	kNN [17]	relevance	CNN ad	0.7906	0.4569	0.5622
	kNN [17]	diversity	CNN all	0.6935	0.4198	0.5053
	Multi-topic	Proposed	pseudo-rel.	CNN all	0.7171	0.4731
Rocchio [39]		relevance	Text all-CNN all	0.8064	0.4471	0.5584
Rocchio [39]		diversity	Text all-CNN all	0.7521	0.4744	0.5704
RFE [40]		relevance	Cred-CNN all	0.7379	0.4571	0.5528
RFE [40]		diversity	Vis all-CNN all	0.7514	0.4616	0.5574
SVM [28]		relevance	Cred-CNN all	0.7607	0.4615	0.557
SVM [28]		diversity	CNN all	<b>0.9036</b>	0.4594	0.5926
Random Forests [1]		relevance	GLRLM	0.8086	0.4024	0.5134
Random Forests [1]		diversity	HOG	0.7457	0.47	0.5532
AdaBoost [61]		relevance	CNN ad	0.7279	0.4641	0.551
AdaBoost [61]		diversity	CNN gen	0.7136	0.4564	0.54
kNN [17]		relevance	CNN ad	0.825	<b>0.4905</b>	<b>0.5976</b>
kNN [17]		diversity	CNN gen	0.7364	0.4475	0.5443
Overall		Proposed	pseudo-rel.	CNN ad	0.7043	<b>0.4708</b>
	Rocchio [39]	relevance	Text all-CNN all	0.7986	0.4469	0.5578
	Rocchio [39]	diversity	Text all-CNN all	0.7644	0.4681	0.568
	RFE [40]	relevance	Vis all-CNN all	0.7629	0.4319	0.5369
	RFE [40]	diversity	Vis all-CNN all	0.745	0.45	0.5474
	SVM [28]	relevance	Cred-CNN all	0.7439	0.451	0.5485
	SVM [28]	diversity	CNN ad	<b>0.9025</b>	0.432	<b>0.5688</b>
	Random Forests [1]	relevance	GLRLM	0.7802	0.3902	0.4999
	Random Forests [1]	diversity	HOG	0.7432	0.4376	0.5316
	AdaBoost [61]	relevance	CNN ad	0.7047	0.4498	0.532
	AdaBoost [61]	diversity	CNN ad	0.7187	0.4393	0.5249
	kNN [17]	relevance	CNN ad	0.8079	0.4638	0.58
	kNN [17]	diversity	CNN all	0.7165	0.4276	0.5208

We conclude that although the method doesn't perform best in terms of relevance — which is somehow expected given the automated feedback assumption — it achieves good performance in terms of diversification and overall it is capable to achieve comparable/better results than other, user based, RF techniques. It has the main advantage that the user ground-truth is simulated automatically and no human presence is required. Moreover, the RF techniques were simulated with the ideal ground truth, which gives maximum performance. However, in real case, human feedback is not always accurate and reliable due to time constraints, which gives more weight to the achieved results.

Second, we compare to the 2014 MediaEval Retrieving Diverse Social Images benchmarking [22] (Div150Cred dataset). The best performance parameter combination in this case is visual all-textual all-credibility-180-0-0.7-cityblock-single. Results are presented in Table 5. On this dataset one can observe that the use of diversified feedback instead of only relevance allows for improvement over the last one. However, regardless the use of actual image ground truth, the best traditional relevance feedback result in terms of  $F1@20$  is 0.5172, achieved with SVM and Radial Basis Function (RBF) kernel. This is around 5 percentage points less than the proposed approach. Best result is also achieved in terms of visual diversification,  $CR@20 = 0.4555$ , which is almost 5 percentage points greater than the best performing RF technique on this dataset, SVM-RBF. In terms of relevance, the proposed method is on the 5th place, which is a good result if we take into account that it is focused mainly on the diversification improvement. As for the previous results, these results are very promising considering the fact that the proposed approach uses automatically generated feedback.

## 5.4 Comparison to other diversification approaches

In this experiment, we compare the results against state-of-the-art diversification approaches from the literature.

First, we compare to the 2015 MediaEval Retrieving Diverse Social Images benchmarking [21] (Div150Multi dataset). For the fairness of the comparison, we present only the approaches that use similar techniques (e.g., relevance feedback, hierarchical clustering), namely: UNED [7] — uses relevance human feedback, which is the input to several Local Logistic Regression models that estimate the probability for a certain image to belong to a relevant set; Recod [8] — uses a geographic-based, face-based, and blur-based pre-filtering

**Table 5** Comparison to relevance feedback approaches on Div150Cred dataset [19] (RBF - Radial Basis Function kernel; best results are represented in bold)

RF approach	Feedback	Descriptor	$P@20$	$CR@20$	$F1@20$
Proposed	pseudo-rel.	Vis all-Text all-Credibility	0.7907	<b>0.4555</b>	<b>0.5707</b>
Rocchio [39]	relevance	CN	<b>0.8549</b>	0.3385	0.4718
Rocchio [39]	diversity	CSD	0.7126	0.3429	0.455
RFE [40]	relevance	CN	0.828	0.3239	0.4526
RFE [40]	diversity	CN	0.787	0.3561	0.4773
SVM RBF [28]	relevance	GLRLM	0.8508	0.369	0.505
SVM RBF [28]	diversity	all visual	0.75	0.4086	0.5172
AdaBoost [61]	relevance	GLRLM	0.8077	0.3666	0.4934
AdaBoost [61]	diversity	LBP	0.7463	0.3779	0.4935

**Table 6** Comparison to diversification approaches from 2015 MediaEval Retrieving Diverse Social Images benchmarking [21] (best results are represented in bold)

Set	Approach	Pre-filtering	Modality	$P@20$	$CR@20$	$F1@20$
Single-topic	proposed	yes	CNN	0.713	<b>0.4714</b>	0.5509
	UNED [7]	no	visual-textual	0.7645	0.4194	0.524
	Recod [8]	yes	credibility	0.763	0.4301	0.539
	MIS [64]	yes	CNN	0.7935	0.4546	0.5595
	ETH-CVL [41]	no	visual-textual	<b>0.842</b>	0.442	<b>0.5674</b>
Multi-topic	proposed	yes	CNN	0.7171	<b>0.4731</b>	<b>0.5533</b>
	UNED [7]	no	visual-textual	<b>0.7886</b>	0.4491	0.5519
	Recod [8]	yes	visual	0.735	0.4221	0.5133
	MIS [64]	yes	CNN	0.7636	0.4354	0.5353
	ETH-CVL [41]	no	visual	0.6829	0.4622	0.5333
Overall	proposed	yes	CNN	0.7043	0.4708	0.5471
	UNED [7]	no	visual-textual	0.7766	0.4344	0.538
	Recod [8]	yes	credibility	0.7198	0.4309	0.5219
	MIS [64]	yes	CNN	<b>0.7784</b>	0.445	<b>0.5473</b>
	ETH-CVL [41]	no	visual	0.6853	<b>0.4724</b>	0.5453

strategies followed by a re-ranking and a k-Medoids-based diversification step; MIS [64] — uses face-based and distance-based filters to remove irrelevant images, followed by a distance-based re-ranking and an Agglomerative Hierarchical Clustering approach for diversification; ETH-CVL [41] — uses relevance ground truth provided for development data to train a generic SVM on CNN features for relevance scoring.

Results are presented in Table 6. Overall, the proposed approach achieves the second best result in terms of  $F1@20$ , after MIS [64] —  $F1@20 = 0.5473$ , with  $F1@20 = 0.5471$ . However, we achieve the best diversification performance on *single – topic* and *multi – topic* sets, with  $CR@20 = 0.4714$ ,  $CR@20 = 0.4731$  respectively. Results show that, overall, methods achieving the highest relevance are not necessarily the ones with the highest diversification, e.g., UNED [7] has  $P@20 = 0.7766$  compared to  $P@20 = 0.7043$  achieved with the proposed approach, but their diversification is lower. In terms of modality, exploiting the CNN information allows for the best performance. The use of pre-filtering techniques do improve performance in certain cases, but is not strictly necessary, e.g., ETH-CVL [41] achieves  $F1@20 = 0.5453$  without any filtering.

**Table 7** Comparison to diversification approaches from 2014 MediaEval Retrieving Diverse Social Images benchmarking [22] (best results are represented in bold)

Approach	Pre-filtering	Modality	$P@20$	$CR@20$	$F1@20$
proposed	yes	Vis all-Text all-Credibility	<b>0.7907</b>	<b>0.4555</b>	<b>0.5707</b>
UNED [9]	no	textual	0.7772	0.4343	0.5502
RECOD [10]	yes	Vis-Cred	0.7598	0.4288	0.5423
MIS [65]	no	visual-textual	0.6732	0.3982	0.4949
PeRCeiVe@UNICT [49]	yes	textual	0.7553	0.3902	0.5063



Second, we compare to the 2014 MediaEval Retrieving Diverse Social Images benchmarking [22] (Div150Cred dataset). For the same reason of fairness of the comparison, we present only the approaches that use similar techniques (e.g., relevance feedback, hierarchical clustering), namely: UNED [9] — uses Hierarchical Agglomerative Clustering to group together the images according to latent topics related to the images, detected with Formal Concept Analysis; Recod [10] — uses a geographic-based, and a face-based pre-filtering strategies followed by a re-ranking and a k-Medoids-based diversification step; MIS [65] — uses a distance-based reranking, followed by an Adaptive Hierarchical Clustering approach for diversification; PeRCeiVe@UNICT [49] — uses a faced-based and a GPS-based pre-filtering to remove irrelevant images, followed by Random Forest predictors to define a dissimilarity measure between images, a Cluster-filtering and a Hierarchical Clustering-based ranking.

Results are presented in Table 7. The proposed approach achieves best result in terms of  $F1@20$ , with  $F1@20 = 0.5707$ , which is around 2 percentage points higher than the method from MediaEval 2014, UNED [9] —  $F1@20 = 0.5502$ . We also achieve best diversification performance with  $CR@20 = 0.4555$  and best relevance with  $P@20 = 0.7907$ .

### 5.5 Visual example

The final experiment consists on discussing some visual examples of the diversification results. We present two examples, first with query 'Christ Church Cathedral' in Missouri

#### Flickr initial ranking



#### Proposed method



**Fig. 7** Visual Example for query 'Christ Church Cathedral' in Missouri from *single – topic* set: first two lines represent first 10 retrieved images from Flickr and last two lines represent the first 10 images refined with the propose approach. The order of the images is from top to bottom and from left to right



from *single – topic* set and the second one with query 'Pingxi Sky Lantern Festival' from *multi – topic* set of the Div150Multi dataset. In both examples, we take the first 10 retrieved images, considering two cases: (i) images are retrieved in their initial order from Flickr, which is the baseline and (ii) images are refined by the proposed approach.

The first example is depicted in Fig. 7. Images are retrieved from left to right and top to bottom, according to their rank. We have in the first case  $P@20 = 0.4$  and  $CR@20 = 0.32$ , which means that relevance and diversification are low. We can also observe in the figure that the number of relevant images is low and consequently diversification is poor. With the proposed approach we obtain more relevant images and we can observe that they are also visually diverse. In this case  $P@20 = 0.8$  and  $CR@20 = 0.44$ , which is a visible improvement over the baseline case.

For the baseline of the second example, images are relevant, but more than half of them are redundant, as it can be observed in Fig. 8. With Flickr initial ranking, precision is  $P@20 = 0.95$  and cluster recall  $CR@20 = 0.2778$ . By refining the results with the proposed method, we obtain  $P@20 = 0.8$  and  $CR@20 = 0.6667$ , which is a significant improvement over the initial ranking. This can also be observed in the figure, where images are more diverse in terms of visual information.

**Flickr initial ranking**



**Proposed method**



**Fig. 8** Visual Example for query 'Pingxi Sky Lantern Festival' from *multi – topic* set: first two lines represent first 10 retrieved images from Flickr and last two lines represent the first 10 images refined with the propose approach. The order of the images is from top to bottom and from left to right

## 6 Conclusions

In this article we addressed the problem of social image search result diversification from the perspective of relevance feedback techniques. We proposed a novel perspective that renders the feedback process completely automatic via pseudo-relevance feedback and considers in priority the diversification, in addition to the relevance of the results. The method operates on top of an existing retrieval system.

Experimental validation on Flickr data (datasets from the 2014 and 2015 MediaEval Retrieving Diverse Social Images tasks) show the potential of this approach. It has superior or at least comparable performance in terms of visual diversification than traditional state-of-the-art relevance feedback methods, but its big advantage is the automatic process of selecting images as feedback, which doesn't require a human annotator. Moreover, the proposed approach achieves similar or better performance than other, more complex, state-of-the-art diversification approaches from the literature. It allows in particular to achieve better diversification of the results. We therefore proved the benefits of the pseudo-relevance assumption in the context of result diversification opening new perspectives for this area of research.

Future work will mainly address exploring more complex diversification scenarios, such as the ones involving more general concept queries where relevance and diversification are relative to each concept in particular. This is a scenario more closer to a general purpose retrieval system. However, the lack of experimentation data is a practical issue for tackling this.

**Acknowledgments** This work has been funded by the Ministry of European Funds through the Financial Agreement POSDRU 187/1.5/S/155420 PROSCIENCE, and by the ESF POSDRU/159/ 1.5/S/132395 InnoRESEARCH programme. We acknowledge also the MediaEval Benchmarking Initiative for Multimedia Evaluation (<http://www.multimediaeval.org>) for providing the data.

## References

1. Bosch A, Zisserman A, Munoz X (2007) Image classification using random forests and ferns. In: Proceedings of IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21, pp 1–8. ISSN 1550-5499
2. Boteanu B, Mironică I, Ionescu B (2015) Hierarchical clustering pseudo-relevance feedback for social image search result diversification. In: Proceedings of 13th International Workshop on Content-Based Multimedia Indexing CBMI 2015, Prague, Czech Republic, 10–12, pp 1–6
3. Cao G, Nie JY, Gao J, Robertson S (2008) Selecting Good Expansion Terms for Pseudo-Relevance Feedback. In: ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, 20–24, pp 243–250
4. Cao L, Ji R, Liu W, Gao Y, Duan LY, Men C (2012) Weakly supervised topic grouping of youtube search results. In: Proceedings of the 19th IEEE International Conference on Image Processing (ICIP), Coronado Springs - Disney World Orlando, FL, USA, pp 2885–2888
5. Carbonell J, Yang RY, Brown RD, Geng Y, Lee D (1997) Translingual information retrieval: a comparative evaluation. In: Proceedings of 15th International Joint Conference on Artificial Intelligence (IJCAI), Aichi, Japan, 23–29, pp 708–715
6. Cai D, He X, Li Z, Ma W-Y, Wen JR (2004) Hierarchical clustering of WWW image search results using visual, textual and link information. In: Proceedings of the 12th Annual ACM International Conference on Multimedia, NY, USA, 10–16, pp 952–959
7. Castellanos A, Benavent X, Serrano AG, De Ves E, Cigarrn J UNED-UV @ retrieving diverse social images task. In: Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 14–15, 2015, CEUR-WS.org, ISSN 1613-0073, 1436, <http://ceur-ws.org/Vol-1436/>

8. Calumby RT, do IBA, Araujo C, Santana VP, Munoz JAV, Penatti OAB, Li LT, Almeida J, Chiachia G, Gonaves MA, Da S. Torres R Recod @ MediaEval 2015: diverse social images retrieval. In: Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 14–15, 2015, CEUR-WS.org, ISSN 1613–0073, 1436, <http://ceur-ws.org/Vol-1436/>
9. Castellanos A, GarcEDA-serrano A, Recuero JMC UNED @ retrieving diverse social images task. In: Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, 16–17, 2014, CEUR-WS.org, ISSN 1613–0073, 1263, <http://ceur-ws.org/Vol-1263/>
10. Calumby RT, Santana VP, Cordeiro FS, Penatti OAB, Li LT, Chiachia G, Da S. Torres R Recod @ MediaEval 2014: diverse social images retrieval. In: Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, 16–17, 2014, CEUR-WS.org, ISSN 1613–0073, 1263, <http://ceur-ws.org/Vol-1263/>
11. Datta R, Joshi D, Li J, Wang JZ (2008) Image Retrieval: ideas, Influences, and Trends of the New Age. *ACM Comput Surv* 40(2):1–60
12. Dang V, Croft WB (2012) Diversity by proportionality: an election-based approach to search result diversification. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Oregon, USA, 12–16, pp 65–74
13. Dang-Nguyen D-T, Piras L, Giacinto G, Boato G, De natale F (2015) A hybrid approach for Retrieving Diverse Social Images of Landmarks. In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME), Torino, Italy, pp 1–6
14. Denis F (1998) PAC learning from positive statistical queries. In: Proceedings of 9th International Conference on Algorithmic Learning Theory, Lecture Notes in Artificial Intelligence, 1501. Springer - Verlag, pp 112–126
15. De Weijer V, Schmid C, Verbeek J, Larlus D (2009) Learning color names for real-world applications. *IEEE Trans Image Process* 18(7):1512–1523
16. Garcia Seco de Herrera A, Kalpathy-Cramer J, Demner Fushman D, Antani S, Müller H (2013) Overview of the ImageCLEF 2013 medical tasks. In: Working Notes of Cross Language Evaluation Forum (CLEF), Spain, pp 23–26
17. Giacinto G (2007) A nearest-neighbor approach to relevance feedback in content-based image retrieval. In: Proceedings of the 6th ACM Conf. on Image and Video Retrieval, Amsterdam, Netherlands, 09–11, pp 456–463
18. Ionescu B, Popescu A, Radu A-L, Müller H (2014) Result diversification in social image retrieval: a benchmarking framework
19. Ionescu B, Popescu A, Lupu M, Gînscă AL, Boteanu B, Müller H (2015) Div150Cred: a social image retrieval result diversification with user tagging credibility dataset. In: ACM Multimedia Systems (MMSys), Oregon, USA, 18–20, pp 207–212
20. Ionescu B, Radu A-L, Menéndez M, Müller H, Popescu A, Loni B (2014) Div400: a social image retrieval result diversification dataset. In: ACM Multimedia Systems (MMSys), Singapore, 19–20, pp 29–34
21. Ionescu B, Gînscă AL, Boteanu B, Popescu A, Lupu M, Müller H Retrieving diverse social images at MediaEval 2015: challenge, dataset and evaluation. In: Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 14–15, 2015, CEUR-WS.org, ISSN 1613–0073, 1436, <http://ceur-ws.org/Vol-1436/>
22. Ionescu B, Popescu A, Lupu M, Gînscă AL, Müller H Retrieving diverse social images at MediaEval 2014: challenge, dataset and evaluation. In: Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, 16–17, 2014, CEUR-WS.org, ISSN 1613–0073, 1263, <http://ceur-ws.org/Vol-1263/>
23. Jones S, Shao L (2013) Content-based retrieval of human actions from realistic video databases. *Inf Sci* 236:56–65
24. Jiang L, Mitamura T, Yu SI, Hauptmann AG (2014) Zero-example Event Search Using Multimodal Pseudo Relevance Feedback. In: Proceedings of International Conference on Multimedia Retrieval, Glasgow, U.K., 01–04, pp 297–304
25. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. *ACM Comput Surv (CSUR)* 31(3):264–323
26. Ksibi A, Ammar AB, Amar CB (2014) Adaptive diversification for tag-based social image retrieval. *Int J Multimed Inf Retr* 3(1):29–39
27. King B (1967) Step-wise clustering procedures. *J Am Stat Assoc* 62(317):86–101
28. Liang S, Sun Z (2008) Sketch retrieval and relevance feedback with biased SVM classification. *Pattern Recogn Lett* 29(12):1733–1741
29. Liu B, Lee WS, Yu PS, Li X (2002) Partially supervised classification of text documents. In: Proceedings of 19th International Conference on Machine Learning, Sydney, Australia, 08–12, pp 387–394
30. Ludwig O, Delgado D, Goncalves V, Nunes U (2009) Trainable Classifier-Fusion Schemes: An Application To Pedestrian Detection. In: Proceedings of the 12th Int. IEEE Conf. on Intelligent Transportation Systems, Missouri, USA, 04–07, pp 1–6

31. Manjunath BS, Ohm JR, Vasudevan VV, Yamada A (2001) Color and texture descriptors. *IEEE Trans Circ Syst Video Technol* 11(6):703–715
32. Over P, Awad G, Michel M, Fiscus J, Sanders G, Kraaij W, Smeaton AF, Queenot G (2013) TRECVID 2013 — An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. TRECVID Workshop, NIST, USA. <http://www-nlpir.nist.gov/projects/tvpubs/tv13.papers/tv13overview.pdf>
33. Ojala T, Pietikinen M, Harwood D (1994) Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, 1, pp 582–585
34. Priyatharshini R, Chitrakala S (2013) Association based image retrieval: a survey. In: *Mobile Communication and Power Engineering*, Springer Communications, Computer and Information Science, 296, pp 17–26
35. Paramita ML, Sanderson M, Clough P (2009) Diversity in photo retrieval: overview of the ImageCLEF photo task 2009. In: *Multilingual Information Access Evaluation II. Multimedia Experiments*, pp 45–59
36. Pedronette DCG, Da S. Torres R, Calumby RT (2014) Using contextual spaces for image re-ranking and rank aggregation. *Multimed Tools Appl* 69(3):689–716
37. Popescu A CEA LIST's participation at MediaEval 2013 placing task. In: *Proceedings of the MediaEval 2014 Workshop*, Barcelona, Spain, 16–17, 2014, CEUR-WS.org, ISSN 1613-0073, 1263, <http://ceur-ws.org/Vol-1263/>
38. Rudinac S, Hanjalic A, Larson MA (2013) Generating visual summaries of geographic areas using Community-Contributed images. *IEEE Trans Multimed* 15(4):921–932
39. Rocchio J (1971) Relevance feedback in information retrieval. In: *The Smart Retrieval System Experiments in Automatic Document Processing*. Prentice Hall, Englewood Cliffs, NJ, USA, pp 313–323
40. Rui Y, Huang T, Chang S-F (1999) Image retrieval: current techniques, promising directions and open issues. *Vis Commun Image Represent* 10(1):39–62
41. Ravindranath SS, Gygli M, Gool LV ETH-CVL @ MediaEval 2015: learning objective functions for improved image retrieval. In: *Proceedings of the MediaEval 2015 Workshop*, Wurzen, Germany, 14–15, 2015, CEUR-WS.org, ISSN 1613-0073, 1436, <http://ceur-ws.org/Vol-1436/>
42. Smeulders AWM, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. *IEEE Trans Pattern Anal Mach Intell* 22(12):1349–1380
43. Sarac MI, Duygulu P (2014) Bilkent-RETINA at retrieving diverse social images task of mediaEval 2014. In: *Proceedings of the MediaEval 2014 Workshop*, Barcelona, Spain. CEUR-WS.org, ISSN 1613-0073, 1263, <http://ceur-ws.org/Vol-1263/>
44. Spyromitros-Xioufis E, Papadopoulos S, Gînscă AL, Popescu A, Kompatsiaris I, Vlahavas I (2015) Improving diversity in image search via supervised relevance scoring. In: *Proceedings of ACM International Conference on Multimedia Retrieval*, Shanghai, China, pp 323–330
45. Soleymani M, Larson M (2010) Crowd-sourcing for affective annotation of video: development of a viewer-reported boredom corpus. *SIGIR Workshop on Crowd-sourcing for Search Evaluation*, Geneva, Switzerland, p 23
46. Stricker M, Orengo M (1995) Similarity of color image. In: *IS&T/SPIE's Symposium on Electronic Imaging: Science & Technology*, pp 381–392
47. Sun A, Bhowmick SS (2009) Image Tag Clarity: in Search of Visual-Representative Tags for Social Images. In: *Proceedings of the 1st SIGMM Workshop on Social Media*, ACM, Beijing, China, 19–24, pp 19–26
48. Sneath PHA, Sokal RR (1973) *Numerical taxonomy*. Freeman, London, UK
49. Spampinato C, Palazzo S PerCeive Lab@UNICT at MediaEval 2014 diverse images: random forests for diversity-based clustering. In: *Proceedings of the MediaEval 2014 Workshop*, Barcelona, Spain, 16–17, 2014, CEUR-WS.org, ISSN 1613-0073, 1263, <http://ceur-ws.org/Vol-1263/>
50. Taneva B, Kacimi M, Weikum G (2010) Gathering and Ranking Photos of Named Entities with High Precision, High Recall, and Diversity. In: *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, New York, USA, 03–06, pp 431–440
51. Tang X (1998) Texture information in Run-Length matrices. *IEEE Trans Image Process* 7(11):1602–1609
52. Van Leuken RH, Garcia L, Olivares X, Van Zwol R (2009) Visual Diversification of Image Search Results. In: *Proceedings of the 18th International Conference on World Wide Web*, Madrid, Spain, 20–24, pp 341–350

53. Vee E, Srivastava U, Shanmugasundaram J, Bhat P, Yahia SA (2008) Efficient computation of Diverse Query Results. In: Proceedings of 24th IEEE International Conference on Data Engineering, Cancun, Mexico, 07-12, pp 228–236
54. Vieira MR, Razente HL, Barioni MCN, Hadjieleftheriou M, Srivastava D, Traina Jr C, Tsotras VJ (2011) On query result diversification. In: Proceedings of IEEE International Conference on Data Engineering, Hannover, Germany, 11-16, pp 1163–1174
55. Viola P, Jones MJ (2004) Robust Real-Time face detection. *Int J Comput Vis* 57(2):137–154
56. Van Brummelen GR (2013) Heavenly Mathematics: The forgotten art of spherical trigonometry. Princeton University
57. Wang XY, Zhang BB, Yang HY (2013) Active SVM-based relevance feedback using multiple classifiers ensemble and features reweighting. *Eng Appl Artif Intell* 26(1):368–381
58. Wu HC, Luk RWP, Wong KF, Kwok KL (2008) Interpreting TFIDF term weights as making relevance decisions. *ACM Trans Inf Syst* 26(3):1–37
59. Ward Jr JH (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58(301):236–244
60. Yang Y, Nie F, Xu D, Luo J, Zhuang Y, Pan Y (2012) A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Trans Pattern Anal Mach Intell* 34(4):723–742
61. Yu J, Lu Y, Xu Y, Sebe N, Tian Q (2007) Integrating relevance feedback in boosting for content-based image retrieval. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hawaii, USA, 15-20, pp 965–968
62. Yan R, Hauptmann A, Jin R (2003) Multimedia search with pseudo-relevance feedback. In: Proceedings of International Conference on Image and Video Retrieval (CIVR), Illinois, USA, 24-25, pp 238–247
63. Zhu X, Goldberg A, Gael JV, Andrzejewski D (2007) Improving diversity in ranking using absorbing random walks, pp 97–104
64. Zaharieva M, Diem L MIS @ retrieving diverse social images task 2015. In: Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 14-15, 2015, CEUR-WS.org, ISSN 1613-0073, 1436, <http://ceur-ws.org/Vol-1436/>
65. Zaharieva M, Schwab P A unified framework for retrieving diverse social images. In: Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, 16-17, 2014, CEURWS. org, ISSN 1613-0073, 1263, <http://ceur-ws.org/Vol-1263/>



**Bogdan Boteanu** received a BSc degree in electronics and telecommunications (2011) from Military Technical Academy of Bucharest (2011) and a MSc degree in Advanced Techniques from University "Politehnica" of Bucharest (2014), Romania. Since 2014, he is a Ph.D. student at the same University. He has been actively involved in the organizing of MediaEval Benchmarking Initiative for Multimedia Evaluation. His research interests cover multimedia information retrieval and machine learning for information mining.





**Ionuț Mironică** holds a Ph.D. degree in electronic engineering and telecommunications from University of Politehnica of Bucharest (UPB). His research interests cover multimedia information retrieval, computer vision and pattern recognition.



**Bogdan Ionescu** received the Ph.D. degree in image/video processing from University Politehnica of Bucharest (UPB), Bucharest, Romania, and University of Savoie, Chambéry, France, respectively. He is currently a Tenured Associate Professor with the Faculty of Electronics, Telecommunications, and Information Technology, UPB, and a Courtesy Faculty Member with University of Savoie. He coordinates the Video Processing Group with the Image Processing and Analysis Laboratory, UPB. He has authored over 100 scientific publications. His research interests include multimedia information retrieval, image/video content processing and analysis, and human-computer interaction. Dr. Ionescu is/was a Reviewer for several journals, such as SPIE Journal of Electronic Imaging, IS&T Journal of Imaging Science and Technology, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE SIGNAL PROCESSING LETTERS, IET Computer Vision, and Multimedia Tools and Applications; a Lead Editor of a book entitled *Advances in Computer Vision and Pattern Recognition* (Springer); a Lead Guest Editor of a special issue with *Image and Vision Computing* (Elsevier); an Associate Editor of *Academy Journal of Multimedia*; and the Technical Program Committee or Conference Committee Chair of several conferences, such as the European Signal Processing Conference, the IEEE Content- Based Multimedia Indexing Conference, the ACM Multimedia Conference, the European Conference on Computer Vision, the European Workshop on Visual Information Processing, and the International Conference on Image Processing. He will be the General Chair of the ACM International Conference on Multimedia Retrieval in 2017 and is a lead Organizer/Co-Organizer of several benchmarking campaigns, such as the MediaEval Retrieving Diverse Social Images and Affect Task: Violent Scenes Detection.