Multimodal Retrieval with Diversification and Relevance Feedback for Tourist Attraction Images

DUC-TIEN DANG-NGUYEN, University of Trento and Dublin City University
LUCA PIRAS and GIORGIO GIACINTO, University of Cagliari
GIULIA BOATO and FRANCESCO G. B. DE NATALE, University of Trento

In this paper, we present a novel framework that can produce a visual description of a tourist attraction by choosing the most diverse pictures from community-contributed datasets, that describe different details of the queried location. The main strength of the proposed approach is its flexibility that permits to filter out non-relevant images, and to obtain a reliable set of diverse and relevant images by first clustering similar images according to their textual descriptions and their visual content, and then extracting images from different clusters according to a measure of user's credibility. Clustering is based on a two-step process where textual descriptions are used first, and the clusters are then refined according to the visual features. The degree of diversification can be further increased by exploiting users' judgments on the results produced by the proposed algorithm through a novel approach, where users not only provide a *relevance* feedback, but also a *diversity* feedback. Experimental results performed on the MediaEval 2015 "Retrieving Diverse Social Images" dataset show that the proposed framework can achieve very good performance both in the case of automatic retrieval of diverse images, and in the case of the exploitation of the users' feedback. The effectiveness of the proposed approach has been also confirmed by a small case study involving a number of real users.

CCS Concepts: •**Information systems → Multimedia and multimodal retrieval; Information retrieval diversity;**

Additional Key Words and Phrases: diversification, tourist attraction images retrieval

ACM Reference format:
Duc-Tien Dang-Nguyen, Luca Piras, Giorgio Giacinto, Giulia Boato, and Francesco G. B. De Natale. 2017. Multimodal Retrieval with Diversification and Relevance Feedback for Tourist Attraction Images. *ACM Trans. Multimedia Comput. Commun. Appl.* 0, 0, Article 00 (May 2017), 23 pages.
DOI: 0000001.0000001

1 INTRODUCTION

Ten years after the rise of social networks and image storage services such as Facebook and Flickr, the number of online pictures has incredibly increased, reaching 1.8 billion image shares per day across Flickr, Snapchat, Instagram, Facebook and Whatsapp. Thus, the need for efficient and effective image retrieval systems has become crucial. However, current images search engines - such as those included in popular search engines like Bing and Google, or those provided with

This work has been partially supported by the Regional Administration of Sardinia, Italy, within the project "Advanced and secure sharing of multimedia data over social networks in the future Internet" (CUP F71J11000690002).
Author's addresses: Duc-Tien Dang-Nguyen, (Current address) Insight Centre for Data Analytics at Dublin City University; Luca Piras and Giorgio Giacinto, Department of Electrical and Electronic Engineering, University of Cagliari; Giulia Boato and Francesco G. B. De Natale, Department of Information Engineering and Computer Science, University of Trento.
 1551-6857/2017/5-ART00 $15.00
DOI: 0000001.0000001

(a) Results from Flickr. (b) Results without relevance feedback. (c) Results with relevance feedback.

Fig. 1. An example of the first 10 results retrieved by Flickr default search for the query "Milano Duomo" and the first 10 results of the proposed approach without and with relevance feedback. Figures (b) and (c) provide a more complete view, comparing with figure (a). Moreover, the results in (c) provide a more comprehensive understanding of the queried location with respect to (b) in terms of diversity. Images with green and orange borders are considered similar images, thus only one of them should be selected.

image sharing services such as Flickr - mainly aim to provide users with *exact* results for the queries, which are basically the visually best matches, and usually contain redundant information.

Diversity has been demonstrated to be a very important aspect in the evaluation of the quality of the results expected by users, as it is related to the need of obtaining a comprehensive and complete view of the query results, by avoiding redundant results [6]. Indeed, diversification of search results allows for better and faster search, gaining knowledge about different perspectives and viewpoints on retrieved information sources.

Recently, the idea of diversification for image search results has been studied by many researchers [26, 52], and some international challenges have been also proposed to address this issue (ImageCLEF [30] and MediaEval Retrieving Diverse Social Images Task [17]).

Following [52], in this paper we focus on the problem of diversification of social images describing a tourist attraction at a given location, where the results are supposed not only to provide a set of relevant images related to a specific attraction, but also to provide complementary views, including different perspectives, various combinations of day and time (e.g., night and day), which may provide a more comprehensive understanding of the queried location. An example is illustrated in Figure 1, where in figures (b) and (c), the landmark is represented by multiple details (diversification) which provide a more complete view comparing with figure (a). Moreover, the results in (c) provide a more comprehensive understanding of the queried location than (b), in terms of diversity where all images in (c) are different from the viewpoints or time, while some images in (b) can be considered as similar images (e.g., the ones with green and orange border).

In this paper, we propose a multi-modal retrieval framework that can fulfill the diversification problem described above either automatically or through the feedback of users. This framework exploits textual and visual features, as well as the *user credibility* information, which mainly represents how well each user assigns tags to uploaded images [13]. Starting from a set of images of a tourist attraction that is retrieved through tag information, the first step of the proposed method is to filter out *non-relevant* pictures, i.e., images taken at the queried location that do not show the attraction in foreground (e.g., close-up pictures of people if front of the attraction), blurred or out of focus images. As a second step, we propose to cluster similar images by constructing a particular clustering feature tree (CF tree) which is firstly built based on textual and secondly refined based on visual information. The reason behind the choice of using textual features in the first phase, and then the visual features in the second phase is to make the visual processing task less expensive, and more likely to yield precise results (see more details in sections 3.2 and 4.3). After this step, all images that are visually similar, and have similar textual information are grouped into the same branch of the tree. Finally, in the third step, we fulfill the diversification strategy. This step can be done either automatically (which will be our baseline, as in the preliminary version of this work

that has been presented in [11]), or through the involvement of the user by a tailored feedback mechanism.

The automatic diversification of images of the query location is attained by an agglomerative clustering algorithm that processes the CF tree generated by the second step, and then representative images from each cluster are selected on the basis of the user credibility information.

Thanks to the flexibility of the CF tree, we propose here novel ways to refine the retrieval results by the feedback mechanism, that is designed to update the structure of the tree. Traditionally, feedback is implemented to increase the *number* of relevant images (i.e., the so-called *relevance feedback*) [38], but not to improve the *diversity* among the retrieved images. To increase the diversity, we asked the user to provide her feedback on the retrieved images by labeling them not only by using the usual *Relevant*, and *Non-relevant* labels, but also by including a third label named *Already seen*, that helps avoiding the retrieval of images that will not bring anything new from a perceptual/semantic viewpoint. We show that the proposed update of the CF tree according to the user's feedback, allows improving both the relevance and the diversity.

The assessment of the proposed framework is performed on the dataset that has been released in the MediaEval 2015 [17] task on"Retrieving Diverse Social Images", which has been specifically developed for the task at hand. It is worth to note that the used dataset has been annotated by experts whose judgment is subjective and prone to personal views, and so the obtained results should be evaluated in this light: although it is not possible to have an objective distinction of what is "diverse" and what is not, the relevance feedback paradigm is able to adapt the search according to the judgment of the involved user. For this reason, in this paper, the proposed approach has been also tested in a small case study involving a number of real users where the relevance feedback assessment is conducted with 38 people from different background and locations.

The details of the proposed framework are organized in the paper as follows: in Section 2 the related work is briefly described; in Section 3 the proposed framework for tourist attraction image diversification is described in details; in Section 4 we present an extensive experimental analysis; finally, in Section 5 some concluding remarks are drawn.

2 RELATED WORK

2.1 Retrieving Diverse Landmark Images

Even if the problem of diversification has been originally addressed within the text-retrieval community [6], more recently several works in the multimedia retrieval community investigated how to retrieve documents that are relevant and fidiversefi enough to provide a more comprehensive and concise answer to the user's query. For instance, in [36] the authors address the visual diversification of image search results with the use of clustering techniques in combination with a dynamic weighting function of visual features to capture the possible several aspects of the queries (e.g., different type of mammal in a set of animal images, or different model of cars in a motor expo). In [14] the authors propose an approach for label propagation, which favors the propagation of an objectfis label to a set of images representing as many different views of that object as possible by using a random forest framework. In [34], the authors propose a method that can group images of the same viewpoints, and thus can create a comprehensive description of a landmark for users. In [2] the problem of diversification has been faced in order to improve the descriptive power of a multimedia social event summarization framework and to generate holistic visualized summary from microblogs with multiple media types. The authors first partition the images within an event into groups via spectral clustering, then, for each group apply a manifold algorithm to identify the top-ranked image as representative.

Another field where diversification gained more and more influence is the social image search of landmark images [4, 5]. The literature in this field considers both *relevance* and *diversity* as two core criteria for efficient landmark image retrieval systems.

Image *relevance* is commonly estimated from textual information, e.g., analyzing image tags [43], and current search engines are still mainly based on this data. However, textual information is often inaccurate, as it is quite common for users to tag an entire collection with only one label. Accordingly, to improve the relevance of the retrieved results, some works exploited low-level image descriptors, such as SIFT [43], different color histograms [44], or a fusion of textual and visual information [20]. However, while low-level visual descriptors help improving performances, nonetheless they often fail to provide high-level understanding of the scene. It is largely recognized that contextual information can provide significant clues for bridging the gap between low level features and high-level understanding. In this respect, there are many works [7, 20] that make use of contextual information (e.g., GPS) combined with low-level features to boost the performance.

Diversity in the set of retrieved images is achieved by clustering similar images based either on textual or visual properties [20]. A criterion to measure the diversity of the results in image retrieval tasks, and a novel attempt to optimize directly this criterion is proposed in [16]. Some approaches use the concept of "canonical view" [40], where an unsupervised learning algorithm is used to diversify the search results. In [3], the authors propose to exploit the visual saliency to re-rank top results and improve diversification.

A different source of information is used in [13], where a measure of *'user credibility'* is introduced to assess the goodness of the image-tag pairs uploaded by users. This information is extracted from a large amount of annotated data and can be integrated with different cues to improve the landmark search performance, as proposed for the first time in the MediaEval 2014 contest. Indeed, the MediaEval Benchmarking Initiative for Multimedia Evaluation organizes since 2013 a task on retrieving diverse social images (http://www.multimediaeval.org/), by publishing a large collection of landmark images with the ground truth annotated by experts.

2.2 Relevance Feedback

The Relevance Feedback (RF) paradigm has been proposed to refine retrieval results in the context of image classification and retrieval tasks, both to overcome inaccuracies in textual information, and to bridge the semantic gap between the low level image descriptors and the user semantics. This paradigm introduce the *human in the loop*, by asking the user to label a set of images as being relevant or not [41] with respect to her interests. In general, the approaches proposed in the literature to exploit the RF paradigm can be divided into two groups. One class of techniques exploit relevance feedback by modifying some parameters of the search, such as the query vector, by computing a new query vector in the feature space [37], or the distance measure, by using a weighted distance [31, 38]. Another group of approaches are based on the formulation of RF in terms of a pattern classification task, by using popular learning algorithms such as SVMs [24], neural networks and self-organizing maps [8, 22, 49].

To reduce the number of images to be returned to the user, some papers in the past years proposed the use of image clustering techniques [41]. These approaches exploit the hierarchical indexing structure of the clusters to refine the number of images to consider [28, 45, 50]. In [21], the authors exploit the users' feedback to modify the centroid of the considered clusters, but not to refine the number and the shape of the clusters. More recently relevance feedback has been used to expand the query for sketch-based retrieval to improve the final result and return more relevant images [33]. Other works, instead, exploit a sort of *implicit* relevance feedback, known as collaborative filtering, for personalized POI recommendations. Differently from the approach proposed in this

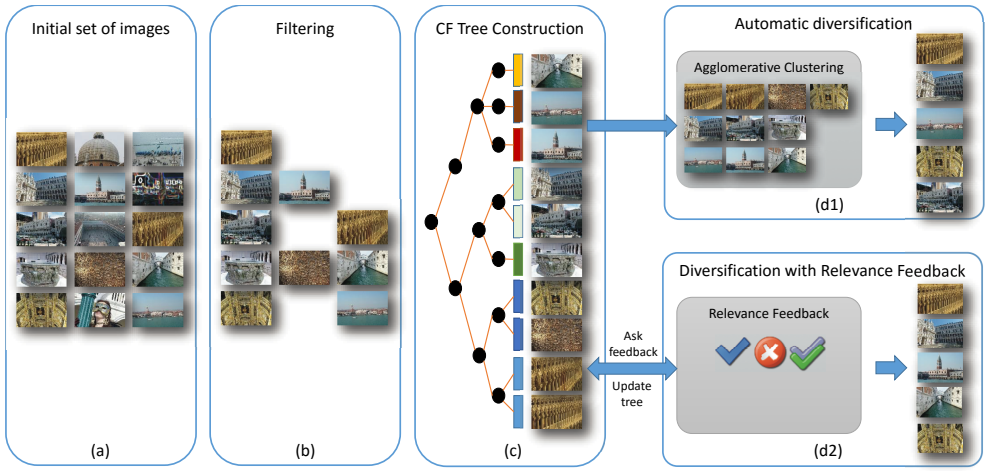

Fig. 2. Schema of the proposed framework.

paper, collaborative filtering is used in [19] to provide recommendations according to similar usersfi visiting history for the creation of an automated travel planning. Likewise, an *implicit* relevance feedback approach has been also used to re-rank the results in image search tasks. In [25] the authors fuse the visual information, social userfis information and image view times to boost the diversity performance of the search result by simply removing the duplicate images from the same user.

Some methods exploit the participation of humans by collecting the feedback from the results to improve the diversification [4]. This method, however is not efficient since it requires feedback on many images (at least 20) for each cluster, and thus requires feedback on at least 400 images for a single location. An improved version of [4] is proposed in [5], where the authors replaced the user RF with the pseudo-RF, by considering the top images from Flickr as relevant, while the ones in the lowest positions as non-relevant. Differently from [4] and [5], the method proposed in this paper introduces a new type of feedback called "*Already seen*" that is aimed to improve not only the diversification, but also the effectiveness of RF.

In this paper, not only we propose to exploit the RF paradigm to improve or modify the indexing structure based on the feedback, but also, as detailed in the next Section 3.4, a new type of feedback has been introduced (that has been named *Already seen*) in order to adapt this paradigm to the diversification problem.

3 METHODOLOGY

The problem can be formulated as follows:

- Let us assume to have a set of touristic attractions $A_i (i = 1 \ldots L)$, each attraction being described by M_i images. For each image, textual, visual and other contextual information is available.
- The task is to select, for each touristic attraction A_i, the subset of N images (where N is a design parameter) that best describes the attraction.

To solve this task we propose to perform a three-steps process, as illustrated in Figure 2. Starting from an initial set of images retrieved from a repository according to the tag information, the first step is to filter out non-relevant images, i.e., those images that have been taken outside of the queried location, or that have been taken in the specific place but the landmark is not in the foreground (e.g., close-up pictures of people), or that are blurred or out of focus. An example is shown in Figure 2, where panel (a) depicts the initial set of images retrieved from the query

"doge_s_palace" using the Flickr default search, and panel (b) depicts the result after filtering (more details in Section 3.1).

Then, to select the subset of N images that best describes the attraction, it is quite straightforward to resort to clustering algorithms that help finding groups of similar images, so that diverse images could be picked form different clusters. Two questions arise:

- Which algorithm to use.
- How to combine the textual, visual, and contextual information that is extracted for each image.

To answer the first question, we observe that, for a given touristic attraction, no assumption can be made about the number of clusters the available images can be grouped into. Consequently, the family of hierarchical clustering algorithms can be chosen for their ability to help finding natural clusters when their number is not a priori known.

To cluster images, we can resort either to the text describing the images, or to the content. We can observe that while the visual content of images of the same attraction can be quite similar even for different views, their textual description usually provides additional information that allows assessing the viewpoint or the detail of the attraction that is in the image. So, we decided to produce a first clustering result using the textual features, and then refine the results by resorting to the visual feature (please see the validation in Section 4.3).

Among the many hierarchical clustering algorithms available, we opted to use the Balanced Iterative Reducing and Clustering (BIRCH) algorithm [51] not only for its speed and accuracy in the results, but also for the feasibility of using the produced clustering feature (CF) tree for further refinement of the clustering result. In particular, the BIRCH algorithm allowed us to build a tree based on textual information first, and then to refine the tree using visual information. An example of the constructed CF tree is shown in panel (c) of Figure 2, where images in the same branch are not only visually similar, but they are also coherent in the textual information (more details in Section 3.2).

Moreover, we could exploit the flexibility of the CF tree produced by the BIRCH algorithm to formulate different approaches for creating a diversified visual description of the landmark. In this work we propose two diversification strategies, namely, in an automatic way or by asking feedback from the user. Automatic diversification is carried out by first applying an agglomerative clustering approach to the CF tree to form the clusters. These clusters are then sorted based on their size, and the image with the highest user credibility is selected from each cluster. In panel (d1) of Figure 2 we show an example of the final result of the process where the queried location is presented by a small set of representative and diverse images (more details are in Section 3.3). By the use of RF to produce a diverse set of images, the structure of the tree is constantly updated until the user is satisfied. Panel (d2) of Figure 2 illustrates the set of images of "doge_s_palace" obtained using the proposed Relevance Feedback paradigm which allows three types of feedback: *Relevant* \ *Non-relevant* \ *Already seen* (more details in Section 3.4).

In both cases, it can be noticed that the final set provides a diversified view of the landmark, with images which are both relevant and represent various viewpoints (e.g., the inside, the outside, day and night pictures, details).

3.1 Filtering outliers

The goal of this step is to filter out 'outliers' by removing images that can be considered as non-relevant. Deriving from the rules in [17], we define an image as non-relevant in the following cases:

(1) It contains people as the main subject. This can be detected by analyzing the proportion of the human face size with respect to the size of the image. In our method, we decided to use

Luxand FaceSDK[1] as face detector, as one of the best face detector in the wild, commonly used for research purposes. A detected face is confirmed as being a human face after checking its color in the H channel (in the HSV color space), since it avoids mis-detection in the case of an artificial face (e.g., a face of a statue).

(2) It was shot far away from the queried location. If an image is geo-tagged, the distance of its GPS location (ϕ, λ) to the queried location (ϕ_l, λ_l) is computed by the Haversine formula: $d_{GPS} = 2R \arcsin\left(\sin^2(\frac{\phi_l - \phi}{2}) + \cos(\phi_l)\cos(\phi)\sin^2(\frac{\lambda_l - \lambda}{2})\right)^{\frac{1}{2}}$, where $R = 6356.752$ km is the Earth radius.

(3) It is out of focus or blurred. An image can be counted as out of focus by estimating its focus. Here, we estimate the focus by computing the absolute sum of the wavelet coefficients and comparing it to a threshold, according to [15].

(4) It received very low number of views, according to the statistics of the social web site where the image has been uploaded. Since we are working on social images datasets (e.g., Flickr), if an image received a low number of views, it can be considered as an outlier because it does not attract the viewers. On the other hand, we would like to stress that if an image received a high number of views, it does not imply that the image is relevant to the query.

After this step, all the remaining images are considered as relevant to the query, and are then passed to the next step.

3.2 Building clustering feature tree by using BIRCH algorithm on textual-visual descriptors

In this step, we use the BIRCH algorithm [51] on textual and visual descriptors to build the CF tree from the filtered set of images.

We opted for the BIRCH algorithm because it allowed us to devise an original way to combine textual and visual information together. BIRCH typically finds a good clustering with a single fast scan of the dataset, then a few additional scans can be performed to further improve the quality of clustering. The goal of the additional scans is to find items that have been assigned to the wrong cluster during the first scan. This phase is computationally less expensive with respect to the first scan as it is performed directly on the sub-clusters in the leaves of the CF tree. In addition, this second scan is less sensitive to the order in which images are considered because the leaf entries of the initial CF tree are structured according to the order of data provided to the first scan.

We exploited this feature of the BIRCH approach by building a tree in the first scan using the textual information only. Then, we performed the refinement of the clustering result by taking into account the visual information only. In other words, the visual similarity between images has been used to refine the clusters that have been formed according to the textual description. The reason behind the choice of using textual features in the first phase, and then the visual features in the second phase is that as soon as the pictures are taken in the same place, their visual similarity is high and confusion may arise. Thus, the number of images to be processed according to the visual description is reduced into smaller more coherent subsets, thus making the visual processing task less expensive, and more likely to yield precise results. Effectiveness of the proposed solution will be demonstrated in Section 4.3. In addition to this very useful feature, BIRCH [51] is also well known to handle effectively noisy data, as in the considered scenario of social images.

BIRCH builds a dendrogram known as a clustering feature tree (CF tree), where similar images are grouped into the same cluster or the same branch of the tree. The whole BIRCH procedure is summarized in Algorithm 1 and for this step, we apply the first two phases. In phase 1, the CF

[1] luxand.com

ALGORITHM 1: Image clustering according to BIRCH

Input: Textual feature vectors X, visual feature vectors V, threshold T, and the branching factor B.
Output: A set of clusters K.
Method: (pseudo-code)

Phase 1 Build an initial CF tree by scanning through the textual feature vectors X with a given T and B.
Phase 2 Update T and rebuild the CF tree based on visual feature vectors V.
Phase 3 Use agglomerative hierarchical clustering (AHC) on CF leaves to form the set of clusters K.

ALGORITHM 2: Building a CF tree

Input: The set of feature vectors X, threshold T, the maximum radius of a cluster R, and the branching factor B.
Output: a CF tree $CFTree$.
Method:

Start an empty tree: $CFTree = \emptyset$
for each $x_i \in X$ **do**
  if $CFTree = \emptyset$ **then**
    Create a new cluster $c = \{x_i\}$ and add c to the root of the $CFTree$.
  else
    Starting from root node, find along the $CFTree$ the closest cluster c to x_i
    if the radius $R(c \cup \{x_i\}) < T$ **then**
      $c = c \cup \{x_i\}$
    else
      Create $c' = \{x_i\}$ and add c' to the father node d of c.
    end if
    Split the nodes if they contain more than B children:
    while the number of children node of $d > B$ **do**
      Split d into d_1 and d_2
      d = father node of d
    end while
  end if
end for

Metrics: Given a cluster $c = \{x_t\}$ where $x_t \in X$ and $t = 1, 2, \ldots, N$, the centroid x_0, and radius $R(c)$ of c is defined as: $x_0 = \frac{\sum_{t=1}^{N} x_i}{N}$, and $R(c) = \left(\frac{\sum_{t=1}^{N}(x_t - x_0)^2}{N}\right)^{1/2}$.

tree is built using the textual feature vectors X. A CF tree is a height-balanced tree which is based on two parameters: the branching factor B, and the threshold T. The CF tree is built by scanning through the descriptors (textual feature vectors X) in an incremental and dynamic way. When each input feature vector is encountered, the CF tree is traversed, starting from the root and choosing the closest node at each level. When the closest leaf cluster is found, a distance between the vector and the candidate cluster is computed. A test is performed to see whether the vector belongs to the candidate cluster or not by comparing the distance with T. If it is smaller than T, the input feature vector is added to that cluster. If not, a new cluster is created and added to the father node. Then, any node that contains more than B children is split. This procedure is summarized as Algorithm 2. The selection of the threshold T is strongly based on the selected features, while changing the value of branching factor B greatly influences the ratio of the width and the height of the tree. In our experiments, the initial values of these parameters are empirically chosen as $T = 0.002$ and $B = 4$.

BIRCH provides an optional phase to "restructure" the tree obtained in the first step in order to obtain a more tidy and compact tree. We have used such a possibility, but we replaced the textual features with the visual features. For each node, its center and radius are recomputed based on the visual feature vectors V instead of the former textual feature vectors X. The value of T is then

updated using the largest radius from leaf clusters (computed on the visual feature distances). Phase 2 of the Algorithm 1 is applied by rebuilding the tree after increasing T and keeping the same value of B. It is worth noting that by setting the value of T as the largest value of the radius from the leaf clusters, the smaller leaf nodes will be merged, and thus the structure of the tree will be updated, while setting the value of T to a small value (e.g., equal to the smallest value of the radius), only some leaf nodes will be split, and thus the structure of the tree will be strongly influenced by only textual information.

As a result of this step, images that are both visually similar and have the similar textual information are grouped into the same branch of the tree.

3.3 Automatic Diversification

To produce diversified results automatically, we start from the CF tree, and obtain the clusters by applying the agglomerative hierarchical clustering (AHC) algorithm [1] (phase 3 of Algorithm 1) on the CF leaves to form the set of clusters. Agglomerative hierarchical clustering is a bottom-up method, starting by considering every leaf as a cluster, then at each iteration the two closest clusters are merged, until the minimum distance is greater than a threshold, or the number of disjoint clusters reaches a limitation threshold. From each cluster, representative images that best describe the queried location are selected. Here, we propose a novel way for choosing such images by exploiting *user credibility* information.

This kind of features were introduced in [13] because in a social image dataset the quality of annotations provided by different users can vary strongly and it was necessary a measure of how figoodfi a user is in tagging. *User credibility* information isestimated by exploiting ImageNet, a manually labelled dataset of 11 million images of around 22, 000 concepts. For each user, at most 300 images which have tags that matched with at least one of the ImageNet concepts are selected. Tags are then analyzed against the corresponding concepts to obtain individual relevance scores. Details on how the scores were computed can be found in [13]. These descriptors are composed by several fields (see Figure 4) and in this work we exploit the *visual score* of a user that, according to the definition given in [17], represents the relevance of the images uploaded by that user, in order to select representative images for each cluster.

To choose the most relevant and diverse images of the landmark, first the clusters are sorted based on the number of images, i.e., clusters containing more images are ranked higher. Then, we extract images from each cluster till the maximum number of required images is reached (e.g., 20 images). In each cluster, the image uploaded by the user who has highest visual score is selected as the first image. If there is more than one image from that user, the image closest to the centroid is selected. If more than one image have to be extracted from a cluster to reach the exact number of images required to build the final set, we select the second image as the one which has the largest distance from the first image, the third image as the one with the largest distance to both the first two images, and so on.

The distance between two images i and j is computed as the visual distance: $d_v(i,j) = ||\overrightarrow{v_i} - \overrightarrow{v_j}||$ where $\overrightarrow{v_i}, \overrightarrow{v_j}$ are the visual descriptors and $||\cdot||$ is the Euclidean distance.

The results in the automatic diversification, as depicted in panel (d1) of Figure 2, will be further analysed in Section 4.3 and considered as a baseline for the RF-based approaches described in the following section.

3.4 Diversification with Relevance Feedback

The main novelty of this paper is to apply RF to diversification, as shown in Figure 2 (panel (d2)). Indeed, thanks to the high flexibility of the proposed framework it is possible to introduce the

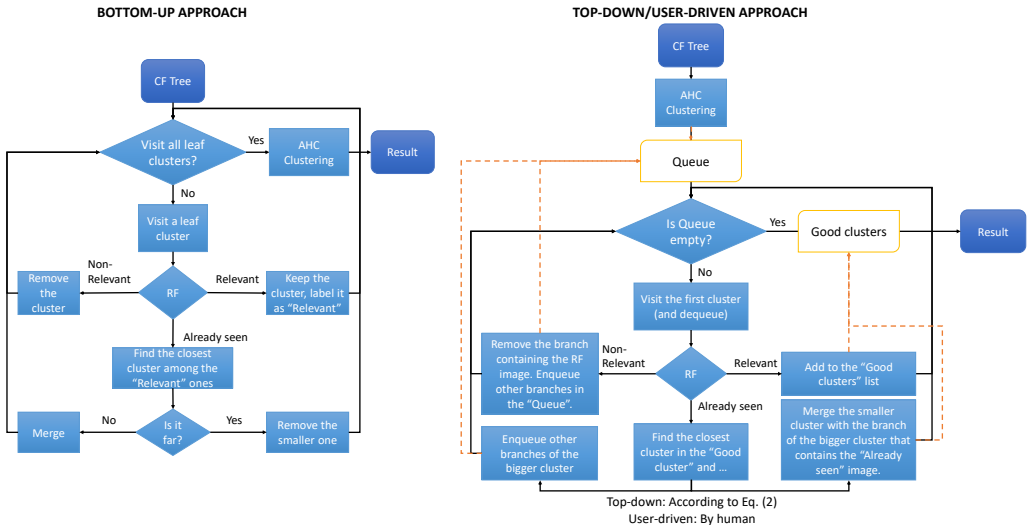

Fig. 3. Flowcharts of the proposed RF approaches.

human in the loop by asking the user to judge the images selected by the system, and, thanks to her feedback, to improve the diversity and the relevance of them. We exploited the flexibility of the proposed framework by formulating three novel Relevance Feedback approaches (summarized in Figure 3), namely a *Bottom-up approach*, a *Top-down approach*, and an *User-driven approach*. In addition, we compare these three novel approaches to the usual RF paradigm used as an additional step after the automatic diversification.

3.4.1 Bottom-up approach. Starting from the CF tree obtained in the second step of the framework (see Section 3.2) the Bottom-up approach goes through all the BIRCH tree from the leaf nodes that are the clusters at the lowest level to the root. This approach selects, from each cluster, a representative image, i.e., the image that is closest to the center based on the visual features distance, and then asks the user to label them as being *Relevant* \ *Non-relevant* to the query or *Already seen*, i.e., relevant to the query but that does not bring anything new from an user's perceptual/semantic viewpoint. After the feedback of the user has been given, the tree is modified according to the following procedure:

- If the representative image has been labeled as *'Relevant'*, the cluster is kept.
- If the representative image has been labeled as *'Non-relevant'*, the cluster is removed.
- If the representative image has been labeled as *'Already seen'* the system finds the closest cluster among those that have been labeled as *'Relevant'*. The distance between two clusters A and B is evaluated according to the following equation:

$$d(A, B) = d_v(a, b) \tag{1}$$

where a and b are the representative images of clusters A and B, respectively.

If the closest cluster is located more than 3 levels away (based on the structure of the tree), i.e., it is too far, the smaller cluster is removed (i.e., it could be a noisy cluster), otherwise, the smaller one is merged with the bigger one and the BIRCH tree is updated accordingly.

After this step, the AHC is applied and the set of images of the queried location is shown.

3.4.2 Top-down approach. This approach does not start from the clusters at the lowest level of the tree, but goes through the BIRCH tree by examining a budgeted number N of big clusters obtained by the agglomerative hierarchical clustering algorithm on CF leaves, similarly to the automatic diversification method. The list of these clusters is called the cluster *'Queue'*. For each cluster in the queue, the representative image is selected, and it is shown to the user to gather her feedback. It is worth noting that this RF method does not need the user credibility information.

- If the representative image is labeled as *'Relevant'*, the cluster is moved to another list, named *'Good Clusters'* list.
- If the representative image is labeled as *'Non-Relevant'*, the branch that contains the non-relevant image is removed and all other branches of that cluster are enqueued to the cluster *'Queue'*.
- If the representative image has been labeled as *'Already Seen'*, the closest cluster in the *'Good Clusters'* list is selected according to Eq. (1). Then, the smaller cluster is kept, while the bigger cluster is split so that the branch that contains the *'Already seen'* image is merged with the smaller cluster, and the other branches are enqueued to the cluster *'Queue'*.

The process stops when the cluster *'Queue'* is empty. Thus, the representative images of the clusters in the *'Good Clusters'* list will be chosen as the ones that best represent the queried location.

3.4.3 User-driven approach. This approach is similar to the *Top-down approach*, but when the user labels an image as *'Already seen'*, instead of evaluating the distances using the clusters belonging to the *'Good Clusters'* and finding the closest cluster, the system allows the users to select the cluster where the image has been already seen. By using this approach it is possible to better exploit the feedback of the user, and to reduce the total number of images for which feedback is required (see Section 4.4).

3.4.4 Refining by common Relevance Feedback. In this approach, the usual dichotomous Relevance Feedback paradigm, that asks the user to assign the labels *Relevant* \ *Non-relevant* to the retrieved images, has been used as an additional step to the proposed framework, just after performing the automatic diversification task. The system asks the user to label the representative images of the top N results returned by the automatic diversification procedure, and the number of images that have been labeled as being *Relevant* \ *Non-relevant* for each cluster is computed. Then, the clusters are sorted as follows:

- Clusters that have a large number of relevant counts are sorted higher.
- Clusters that have the same number of relevant counts are sorted based on the number of non-relevant counts (i.e., a cluster that contains a larger number of 'non-relevant' images should be selected later).
- Clusters that have the same number of *Relevant* \ *Non-relevant* counts are sorted on the basis of the number of images.

For each cluster, the images that are selected to represent the queried location are chosen in the same way as in the automatic diversification step described in Section 3.3.

4 EXPERIMENTAL RESULTS

4.1 Data and Evaluation Metrics

In order to evaluate the proposed method, we ran the experiments on the public dataset MediaEval 2015 "Retrieving Diverse Social Images" [17]. This dataset is built from around 86,000 images from over 300 locations spread over 35 countries allover the world. The images were collected from Flickr by submitting queries on the location names through the Flickr standard interface. For each

```
<credibilityDescriptors>
  <visualScore>0.791</visualScore>
  <faceProportion>0.013</faceProportion>
  <uploadFrequency>395.919</uploadFrequency>
  ...
</credibilityDescriptors>
<photos>
  <photo date_taken="2013-08-19 14:11:49"
  id="9659825826"
  latitude="42.36115" longitude="-71.03523"
  tags="boston nhl massachusetts suffolkcounty nationalhistoriclandmark
  unitedstateslightshipnantucketlv112 lightshipno112"
  title="United States Lightship Nantucket (LV-112)"
  userid="21953562@N07"
  views="533" />
  ...
</photos>
```

Fig. 4. Example of the metadata provided by MediaEval 2015 "Retrieving Diverse Social Images" task.

image, the Flickr metadata (e.g., image title, image description, image ID, tags) are also provided together with the content descriptors that consist of visual, textual and user credibility information. An example of metadata is reported in Figure 4. The images are annotated with respect to both relevance and diversity by experts with advanced knowledge of the locations. The ground truth for the dataset was created by grouping relevant images into different clusters, where each cluster depicts an aspect of the queried location (e.g., side-view, close-up view, drawing, sketch). The dataset is split into two sets: the developing set (devset), and the testing set (testset). The devset consists of 153 one-concept location queries, each location containing 20-25 clusters. The testset contains the results of 139 queries: 69 one-concept location queries and 70 multi-concept queries related to events and states associated with locations, each query containing 2-25 clusters. On average, each query contains 20 clusters. It is worth noting that the devset is made up of the whole dataset of the previous MediaEval 2014 "Retrieving Diverse Social Images" task [18].

The following standard metrics are used to assess the performance with respect to relevance and diversity:

Precision. The relevance is assessed by measuring the precision at N ($P@N$), defined as:

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

where N_r is the number of relevant images in the first N ranked results.

Cluster recall. The diversity is assessed by measuring the cluster recall at N ($CR@N$), defined as:

$$CR@N = \frac{N_c}{N_{tc}} \tag{3}$$

where N_c is the number of clusters found in the first N ranked results and N_{tc} is the total number of clusters of the queried location.

Finally, to assess both relevance and diversity, the harmonic mean $F1@N$ of $P@N$ and $CR@N$ is considered:

$$F1@N = 2 \cdot \frac{P@N \cdot CR@N}{P@N + CR@N} \tag{4}$$

In the reported experiments, all the above measures are considered with different values of the cut off point, namely $N = 5, 10, 20, 30, 40, 50$. It is worth noting that as for many queries the results are grouped in more than 20 clusters, then, according to Eq. (4), the MediaEval benchmarking metric $F1@20$ will be always lower than 1.

Beside relevance and diversity, we also evaluate the RF approaches based on the number of feedback images f, which represents how heavy the interaction with users.

4.2 Feature descriptors

Although the proposed method can be used with any kind of visual descriptors, the choice of the descriptors could influence the results and should be adapted to the specificity of the data.

According to our experiments, the best performances were obtained by using the following visual descriptors:

- Global color naming histogram (CN): maps colors to 11 universal color names: black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow [12].
- Global Color Structure Descriptor (CSD): represents the MPEG-7 Color Structure Descriptor computed on the HMMD color space [27].
- Histogram of Oriented Gradients 2×2 (HOG2x2): Each descriptor consists of 124 feature values, obtained by stacking 2×2 neighboring HOG descriptors each consisting of 31 dimensions [9]. The descriptors extracted from the training images are clustered using the k-means algorithm to identify 300 representative centroids (one per cluster). A histogram of 300 bins is computed from each image, each bin representing the number of image's descriptors assigned to the corresponding centroid. A number of additional histograms are computed using the same procedure, respectively splitting the image into 2×2 and 4×4 blocks, eventually yielding a total of 21 histograms per image (i.e., $21 \times 300 = 6,300$ features) [46].
- Dense SIFT: SIFT descriptors are densely extracted [46] using a flat rather than Gaussian window at two scales (4 and 8 pixel radii) on a regular grid at steps of 5 pixels. The three descriptors are stacked together for each HSV color channels, and quantized into 300 visual words by k-means, and spatial pyramid histograms are used as kernels [23].
- Global Locally Binary Patterns computed on gray scale representation of the image (LBP) [29].

In all the reported experimental results, we used the above mentioned visual descriptors and concatenated them to have the final visual features vector for each image. On the other hand, textual descriptors have been represented by the usual TF-IDF measure, provided by the organizers of the MediaEval 2015 "Retrieving Diverse Social Images" task.

As mentioned in Section 3, the field of the user credibility descriptors that has been used in this work is the *visual score*, which was obtained through visual mining over 17,000 ImageNet visual models, and whose values were provided by the organizers of the MediaEval 2015 "Retrieving Diverse Social Images" task [17]. This score is normalized between 0 and 1, and gives a prediction of how 'good' a user is in tagging, the better the predictions are, the more relevant a user's images should be. More details on this score can be found in [13].

4.3 Evaluation of the automatic diversification method

In this section, we describe the performances attained by the automatic diversification method described in Section 3.3, by also reporting and analyzing the performances attained at each step of the proposed procedure. We used all 153 locations in the devset (which are the whole dataset of MediaEval 2014 "Retrieving Diverse Social Images" [18]) for these experiments.

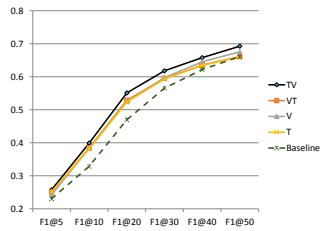

Fig. 5. Clustering step evaluation.

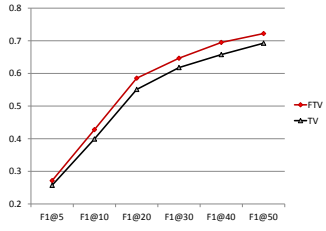

Fig. 6. Filtering step evaluation.

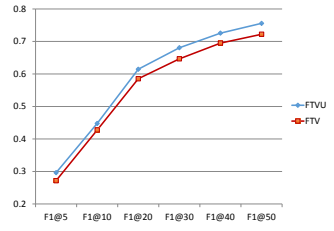

Fig. 7. Diversification step evaluation.

Evaluation of the proposed clustering procedure

In order to evaluate the effectiveness of the proposed clustering step, i.e., using the textual descriptors in the first phase of the BIRCH algorithm, and the visual descriptors in the second phase, we performed a test using four different configurations as follows: i) using visual descriptors only (denoted as V), ii) using text descriptors only (denoted as T), iii) clustering based on visual descriptors and then refined based on textual descriptors (denoted as VT), iv) clustering based on textual descriptors first, and then refined based on visual descriptors (denoted as TV), i.e., the proposed configuration. In this test, the filtering step was not applied, and the centroids of the clusters were selected as representative images, i.e., without using the user credibility information. The performance of these configurations, compared with the 'base-line' using the top N images of the initial set, are shown in Figure 5, where it can be easily seen that at all cut off points, the TV configuration outperforms the others, supporting the considerations in Section 3.2 on using textual information for creating the initial CF tree, and then refining the clusters according to the visual information.

Evaluation of the proposed filtering procedure

The next experiment was performed to evaluate the filtering step. We tested the 4 criteria mentioned in Section 3.1 with different thresholds. The best performance was obtained at 99% true negative rate, i.e., non-relevant images that were correctly classified as non-relevant; 41% of non-relevant images (over all non-relevant images) were detected at that stage. The thresholds used are: (i) the face size is bigger than 10% with respect to the size of the image, (ii) images that were shot farther than 15 kms, (iii) images that have less than 20 views, and (iv) images that have the f-focus value (at the first stage) smaller than 20. In Table 1, we summarize all parameters used in the proposed method.

This is to underline that, although there is still a huge number of non-relevant images in the set, the filter is able to effectively model the "non-relevant class", thus improving the performance of the method. Figure 6 shows the $F1@N$ values of the proposed method with and without the filtering step (denoted as FTV and TV, respectively). In these experiments, clustering is performed by employing the TV configuration, as explained in the previous subsection.

Evaluation of the Diversification step

The importance of user credibility information was assessed by running the best configuration, according to the results in the above tests, with and without the visual score information (FTVU and FTV, respectively). Figure 7 shows the values of $F1@N$ at different cut off points, showing that user credibility information allows improving the performance of the proposed method by attaining a better diversification of the queried location.

Table 1. Parameters used in the proposed method.

Parameter	Goal	Ranges	Empirical values
Branching factor B	Control the ratio of the width and the height of the CF tree.	Equal or bigger than 2. If B is too small, the tree will be very tall (searching on the tree will take more computations); if B is too big, each node will have too many children (reducing the time for searching, but increasing the confusion between tree branches). Thus, selecting the right value for B depends on the data as well as the goals of the retrieval.	With the tourist attraction images, the similarity of the visual information are quite high, thus B should not be very large (i.e., should be smaller than 10). In our experiments, we observed that $B = 4$ gives the best performance.
Tree threshold T	Determine the size (feature space) of cluster feature.	From 0.0 to 1.0. If T is too small, e.g., $T = 0$, almost every image will be a cluster; if T is too big, all images will be grouped into a single cluster. Setting the value of T also depends on the chosen distance metrics as well as the feature vectors.	In our experiment, for the visual information, we empirically set $T = 0.002$.
Face proportion	Filter out images containing people as main subject.	From 0.0 to 1.0. It is easy to verify that a face can be considered large if it is over 10% of the area of the image, while if it is smaller than 5%, it should not be considered as the main subject.	0.1
Distance threshold	Remove images shot far away.	Non negative number.	We want to filter out images that are surely "non-relevant", thus we used a very large value for this threshold: 15km.
Number of views threshold	Filter out non-interesting images.	Non negative number.	Setting this value depends on the application. Since we are working on tourist attraction images, a relevant image has usually a high number of views, thus we empirically select 20 views as the threshold.
F-focus threshold	Detect blurred images.	From 0 to 99: the bigger value the sharper image.	We checked several images and found that the normal quality has the f-focus value of 30 or higher. For values below 30, most of the images are blurred. To ensure that filtered out images are "non-relevant", we set this threshold to 20.

The detailed results for all configurations at cut off points $N = 10, 20, 30$ are reported in Table 2, where it can be easily seen that the configuration labeled as FTVU provides the best performance in terms of all the metrics that have been considered.

4.4 Evaluation of the Relevance Feedback step

In this set of experiments, we used the ground-truth from the MediaEval 2015 "Retrieving Diverse Social Images" task to simulate the behavior of a user who is always consistent with her choice following the rule of the task. This approach allows a fast and extensive simulation which is necessary to evaluate different methods and parameter settings [4]. Such kind of setup represents a common practice in evaluating relevance feedback scenarios [37, 38, 41] and previous experiments from the literature show that results for a small number of iteration are very close to real live user feedback [42].

The aim of all the experiments in this section is to evaluate the number of images for which feedback is required until the user gets 20 images that are both relevant and diverse. We targeted

Table 2. Automatic diversification results at different cut off points.

	P@10	P@20	P@30	CR@10	CR@20	CR@30	F1@10	F1@20	F1@30
FTVU	0.878	0.865	0.823	0.299	0.473	0.588	0.448	0.615	0.681
FVTU	0.859	0.849	0.809	0.296	0.465	0.544	0.440	0.601	0.660
FVU	0.866	0.851	0.819	0.298	0.469	0.559	0.436	0.597	0.655
FTU	0.853	0.833	0.795	0.265	0.424	0.524	0.404	0.562	0.632
FTV	0.854	0.845	0.818	0.268	0.436	0.534	0.407	0.575	0.647
FVT	0.856	0.847	0.813	0.268	0.431	0.530	0.409	0.571	0.642
FV	0.854	0.846	0.811	0.285	0.447	0.541	0.428	0.585	0.649
FT	0.850	0.832	0.795	0.263	0.423	0.524	0.401	0.561	0.632
TV	0.767	0.756	0.752	0.274	0.444	0.539	0.399	0.551	0.618
VT	0.727	0.723	0.716	0.265	0.425	0.521	0.385	0.529	0.595
V	0.769	0.718	0.714	0.260	0.424	0.520	0.384	0.525	0.597
T	0.728	0.716	0.723	0.256	0.415	0.527	0.381	0.523	0.593
Baseline	0.809	0.807	0.803	0.211	0.343	0.450	0.329	0.470	0.565

this goal because the ground-truth of the testset contains 20 - 24 clusters per each location and $N = 20$ is also the official cut off point that has been set in the mentioned competition.

Refining by standard Relevance Feedback

We tested this RF algorithm starting from the top 20 results ($N = 20$) returned by the automatic diversification strategy and asking the user for *Relevant* \ *Non-relevant* feedback. Relevant images that have been already seen were considered as *Relevant*. This setup is named as **RF1**. *Non-relevant* images were removed from the CF tree, and the clusters were resorted (see Section 3.4), so that a new set of top 20 images is returned. The loop was terminated when all the top 20 images returned to the user are labeled as relevant. The tests carried out on all the 153 locations showed that, on average, the algorithm converged after **3.8 iterations** (i.e., the user provides feedback for approximately $\bar{f} = 76$ images). The F1-score at cut off point 20 ($F1@20$) is, on average, equal to 0.676. This is a very interesting result as, after a limited number of iterations, the proposed framework allows exploiting the information provided by the user, thus improving the best result obtained after the automatic diversification step (see $F1@20$ values in Table 2). It is worth noting that the loop was terminated when all the top 20 images returned to the user are labeled as relevant by not taking into account whether the images belong to different clusters or not, so it is possible that the images do not belong to 20 different clusters. That is the reason why the cluster recall $F1@20$ could not reach the maximum value (i.e., 0.876).

In order to better investigate the behavior of the framework in the usual dichotomous Relevance Feedback setup, another experiment, named as **RF2**, was carried out, where relevant images that were already seen have been labeled as *Non-relevant*. In this case, the value of the F1-score reached its maximum value ($F1@20 = 0.876$) because all the top 20 images are relevant, and belong to different clusters. However, this procedure takes on average **13.26 iterations** (approximately feedback is provided for $\bar{f} = 265$ images), i.e., it requires a heavy interaction with users, that is like she is searching for few representative images by examining a very large portion of all the images at the queried location. It is worth to note that usually the Relevance Feedback paradigm in the Content Based Image Retrieval field (CBIR) requires 5 or 6 iteration in each of which the user is asked for feedback on 20 or 25 images [42], so if we compare 256 feedback images with the usual behavior we can observe that we are far beyond the usual range.

The performance attained by the standard Relevance Feedback paradigm has been used as a baseline for assessing the effectiveness of the three novel Relevance Feedback approaches described in Section 3.4 where three different types of feedback were asked for, namely *Relevant* \ *Non-relevant*

Table 3. Results of the Relevance Feedback experiments on 153 locations.

Method	Feedback	F1@20	Avg. # of feedback ($\bar{f}$)
RF1	Relevant = Relevant+Already seen \ Non-relevant	0.676	76
RF2	Relevant \ Non-relevant = Already seen+Non-relevant	0.876	265
Bottom-up	Relevant \ Already seen \ Non-relevant	0.876	102
Top-down	Relevant \ Already seen \ Non-relevant	0.876	92
User-driven	Relevant \ Already seen \ Non-relevant	0.876	49

\ *Already seen*, and the 'user' is simulated according to the ground-truth. In each test, we looped the simulation until reaching the top 20 images, i.e., until all 20 images belong to different clusters, and the F1-score reached its maximum value. We decided to use this setup because the maximum value of the F1-score means that the relevance of the selected images reaches its maximum and all pictures belong to different clusters, and this can well simulate a fully satisfied user.

Bottom-up approach
Tested on all 153 locations, the average number of *Relevant* \ *Non-relevant* \ *Already seen* feedback ($\bar{f}$) are 20, 27.3, and 54.71, respectively. This means that, on average, feedback on 102 images was asked. From Table 3 it is possible to see how in this case we are able to obtain the same maximum value of the F1-score (i.e., 0.876) as the **RF2** approach but with a very small number of feedback images.

Top-down approach
Starting from $N = 15$ branches of the tree (not from the leaf nodes), on 153 locations, the numbers of *Relevant* \ *Non-relevant* \ *Already seen* feedback images are 20, 23.13, and 48.56, respectively, i.e., in total feedback on 92 images was asked.

We would like to stress that the proposed RF model allows limiting the number of feedback images. Thus, we also made a further experiment where the number of feedback images has been limited to 80, and the performance obtained was $P@20 = 0.94$, $CR@20 = 0.68$, and $F1@20 = 0.79$ (the maximum value $F1@20$ is 0.876). Moreover, by limiting the number of feedback images to 76, i.e., approximatively 3.8 feedback iterations, we obtained $F1@20 = 0.75$. This result is quite remarkable, as it shows that we can control the total number of feedback images, and the proposed framework allows adapting the output to the users' goal, thus improving the performance obtained both by the automatic diversification, and by the classical dichotomous RF.

User-driven approach
By allowing a more explicit interaction feedback, by asking the user to also select the cluster where the image has been already seen, and adopting the same setup described above, the number of feedback images significantly reduced to 20, 12.6, and 16.78 for *Relevant* \ *Non-relevant* \ *Already seen*, respectively, thus confirming that the proposed RF paradigm can reduce the number of images for which feedback is required.

The results of the above experiments are summarized in Table 3 where it is possible to notice that the novel RF paradigm described in the three proposed approaches can fulfill different goals of the users. All results are very good in terms of relevance and diversity of the final set of images. Moreover, depending on the application scenario, the target users and the effort they can put in the interaction, the proposed framework allow deciding which approach to take in order to gain high F score values.

Table 4. Results of the Relevance Feedback experiments on 70 multi-concept queries.

Method	Average number of feedback images ($\bar{f}$)
RF1	78
RF2	252
Bottom-up	101
Top-down	82
User-driven	45

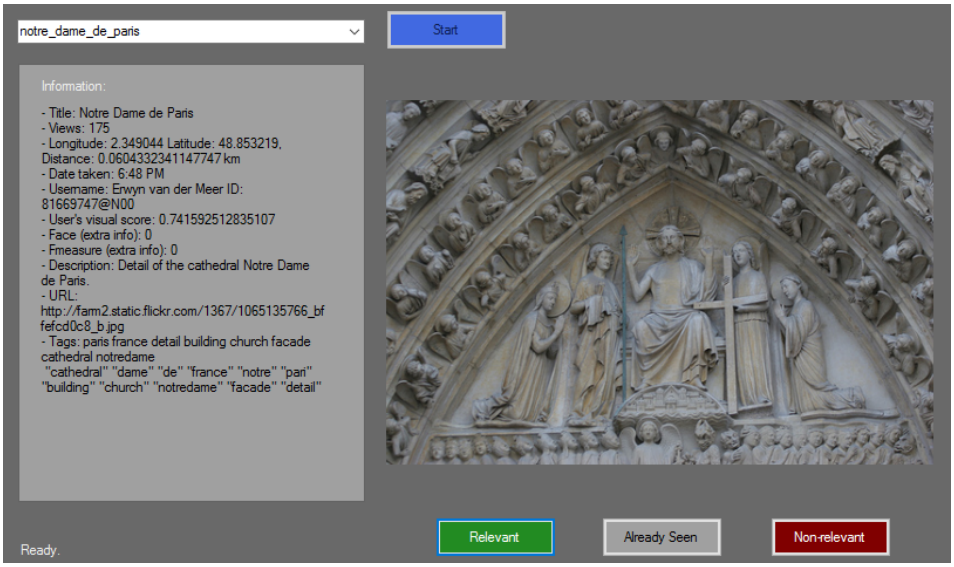

Fig. 8. A screenshot of the application used in gathering feedback from real users.

Experiments on Multi-concept queries

Previous experiments were performed on only one-concept location queries (devset), thus we conduct the next experiments to confirm that the designed RF methods are not only good in the scenario of single location queries, but also on other scenarios like the multi-concept queries (e.g., "Oktoberfest in Munich", "Winter Carnival in Quebec", "Harbin Ice & Snow Sculpture Festival") in the MediaEval 2015 "Retrieving Diverse Social Images" testset. Table 4 shows the average number of feedback images for each RF method, thus confirming the performance of the proposed RF approaches.

Experiments with real users

In these experiments, we tested the user-driven RF method on the first 30 locations with 38 real users: 5 of them are core-users, who performed all 30 locations, and 33 of them are volunteers, who can choose locations from the list and perform the task. An example of our interface is shown in Figure 8. In total, we collected 279 sets of feedback, then we removed 55 sets which were incomplete (i.e., the volunteers did not provide feedback on all images, or selected all images as being relevant in less than 1 minute), to obtain the final 224 sets of feedback. On average 43 feedback images for each location were necessary, which is quite close to the result achieved in the simulated experiments. Table 5 shows the number of feedback images from a core-user using our system on the first 24 locations in the MediaEval 2015 "Retrieving Diverse Social Images" task. It is worth noting that, in this case, some constraints due to real live user feedback experience should be taken into account

Table 5. The number of feedback f from a real user on the first 24 locations in MediaEval 2015 "Retrieving Diverse Social Images" task.

Location	Relevant	Non-relevant	Already seen	Location	Relevant	Non-relevant	Already seen
Angel of the north	20	25	3	Arc De Triomphe	20	37	6
Big Ben	20	20	7	Aztec Ruins	20	15	7
Hearst castle	20	6	8	Berlin Cathedral	20	18	10
Pont Alexander III	20	0	17	Bok Tower Gardens	20	22	4
Neues Museum	20	14	13	Brandenburg Gate	20	15	15
CN Tower	20	26	3	Casa Batlo	20	1	11
Acropolis Athens	20	28	8	Casa Rosada	20	11	4
Agra Fort	20	2	10	Castillo de san Marcos	20	18	10
Albert Memorial	20	23	6	Chartres Cathedral	20	3	4
Altes Museum	20	5	8	Chicken Itza	20	8	10
Amiens Cathedral	20	3	13	Cologne Catedral	20	10	10
Angkor Wat	20	12	5	Colosseum	20	13	6

(e.g., user fatigue, the influence of inter-user agreement). In addition, users have not been provided with any definition of diversity, so they are free to decide if an image is relevant/non-relevant or already seen, and thus the ratio of *Non-relevant* \ *Already seen* is different from the reported results related to user simulations. RF is meant to bring user subjectivity in the system, therefore, it is not negative that the result may change for different users. Indeed, we had different users selecting different set of images but with the same level of measured diversity. Thus, in our tests subjectivity impacts on the results but not on the objective measure.

4.5 Comparison with state-of-the-art methods

In this experiment, we compare our results with those of two state-of-the-art RF approaches from the literature, namely Pseudo-RF [5] and SVM RBF [24]. Table 6 shows the results of Pseudo-RF and SVM RBF on 153 locations in the devset (as reported in [5]), compared to the proposed RF methods.

In Pseudo-RF, the authors used the top and last images in the ranking of the default Flickr search results as the pseudo feedback for relevant and non-relevant images, respectively. The retrieved images are then clustered using a Hierarchical Clustering (HC) scheme following a "bottom up" (agglomerative) approach and finally the results are obtained by selecting images based on the number of the relevant and non-relevant images within each cluster. In that work, different kind of visual, textual and *user credibility* descriptors were experimented individually or in combination. Surprisingly, the "histogram representations of term frequency (TF)" alone proved to be very efficient for diversification, while maintaining a good performance in terms of relevance (see [5] for further details). An explanation for this result may be that the *early fusion* approach used to carry out the combination of the features in that work, probably does not fit properly to a diversification task [32]. In that paper, the authors do not provide a comparison of the perfromances attained by their approach with or without the contribution of the Pseudo-Relevance Feedback, but it is worth to note that even using a significant number of feedback images (i.e., 110 relevant and 18 non-relevant images), their best results measured both as P@20, CR@20, and $F1$@20 are always lower than the ones simply obtained with the **RF1** approach, that required a lower number of feedback images.

In SVM RBF, RF is formulated as a two-class classification task, and, according to the experimental setup used in [5], user relevance feedback is simulated with the imagesfi ground truth in a window of 20 images. These images are chosen among the ones that are furthest from the decision hyperplane in the feature space, and that lie on the side of the hyperplane where the *decision function* is positive. The values of the *decision function* can be used as a measure of the relevance of the images, so that the ones with the highest relevance are most likely to be targeted by the user as the ones she is most interested in, and can be regarded as feedback results, and returned to the

Table 6. Comparisons between the Relevance Feedback approaches on 153 locations.

RF Method	P@20	CR@20	F1@20
SVM RBF [24]	0.851	0.369	0.505
Pseudo-RF [5]	0.819	0.475	0.595
RF1	1	0.511	0.676
RF2, Bottom-up, Top-down, User-driven	1	0.780	0.876

user (see [24] for further details). The Radial Basis Function (RBF) kernel has been chosen for the SVM settings, and according to the preliminary experiment performed in [5], only visual features have been used in this comparison. SVM RBF with respect to Pseudo-RF achieves better results in terms of *P*@20 but not in terms of *CR*@20, and this result clearly shows that methods achieving higher precision, are not necessarily the ones with a higher diversification. This probably is due to the intrinsic nature of the SVM approach more prone to the strict binary classification than to a diversification task

Comparison with RF methods. According to the results reported in Table 6, it easy to see that the proposed RFs outperform the other methods in both precision and cluster recall. It is worth to note that thanks to the flexibility of the proposed framework, the novel ways to refine the retrieval results by the feedback mechanism, allow improving both the precision and the diversity, not only providing a 'diversity' feedback (*Relevant* \ *Non-relevant* \ *Already seen*) but also with the usual dichotomous Relevance Feedback paradigm (**RF1** and **RF2**).

Comparison with state-of-the-art methods. As a final result, we report the comparison with the five methods that achieved the best performance in the MediaEval 2015 "Retrieving Diverse Social Smages" competition, namely TUW [39], USEMP [47], PRa-MM [10], MIS [48], and ETH-CVL [35]. Following the rules of the competition, we tuned the parameters and configurations using the images in the devset, and then applied the methods to the testset. One of the compared methods was our preliminary study submitted to MediaEval 2015: the PRa-MM [10]. Different from that work, in the proposed FTVU, the filtering parameters and the visual descriptors selection were optimized. In TUW [39], the authors first removed the non-relevant images, then clustered them using the K-mean algorithm. Finally, they used two fusion methods, namely a weighted linear algorithm, and Bayesian inference, to re-rank the results. USEMP [47] used a supervised Maximal Marginal Relevance (sMMR) approach, by training the reference model based on relevant and non-relevant examples from other queries. After optimizing the model, an L2-regularized Logistic Regression classifier was used to retrieve the images. In MIS [48], first non-relevant images were filtered out, then the provided user-generated textual descriptions and the visual content of the images were exploited by agglomerative hierarchical clustering. ETH-CVL [35] proposed a linear combination method to quantify how relevant and representative a selected subset is based on several submodular functions: visual representativeness, visual relevance, text relevance, Flickr rank, and time representativeness.

Figure 9 depicts the $F1@N$ measure at all the cut off points, showing that the proposed automatic diversification (FTVU) outperforms all other methods at all cut off points.

Focusing on cut off point $N = 20$ we report more detailed results in Figure 10 for all the variants proposed, including the Relevance Feedback based methods. We would like to stress here again that the ground-truth contains in average 20 clusters per each query, so $N = 20$ is the more "reasonable

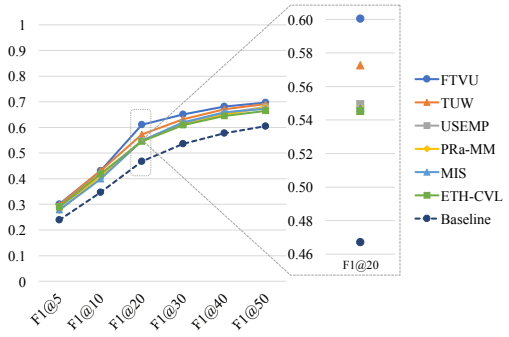

Fig. 9. Performance of FTVU compared to state-of-the-art methods in terms of $F1@N$ for different values of N.

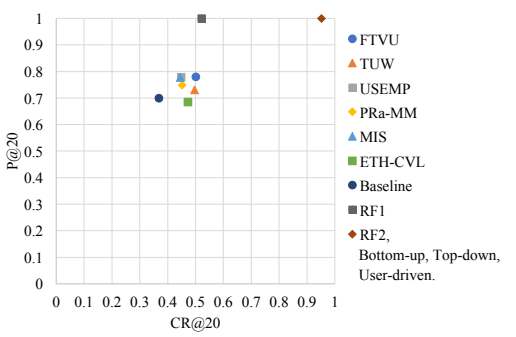

Fig. 10. Performance of FTVU compared to state-of-the-art methods in terms of $P@20$ and $CR@20$.

target" for the RF methods. The proposed framework (both applying automatic diversification and exploiting RF) provides better performances on both $P@20$ and $CR@20$ with respect to state-of-the-art methods. It is possible to see that the idea of including the user in the loop allows increasing both relevance and diversity, thus demonstrating that RF paradigm can be effectively applied also in this field.

4.6 Computational Complexity

Concerning the complexity of the method, we provide here the analysis of the two more impactful steps, clustering and diversification (features extraction can be performed off-line). Both steps are computed based on the CF tree and the construction of such tree requires $O(M)$ comparisons (as mentioned in the original BIRCH study [51]), where M is the number of images, and each comparison is computed as the Euclidean distance between two feature vectors. Let $|X| = dim_X$ be the number of dimensions of the feature vector X (see Algorithm 2), we can rewrite the complexity of the construction step as $O(|X|M)$. For the clustering step, agglomerative hierarchical clustering costs $O(M^2)$ comparisons, and thus the complexity is $O(|X|M^2)$. For the diversification step the complexity is bounded by $f \cdot O(M)$, where f is the number of feedback, since updating the tree in BIRCH requires only $O(log_B(M))$ operations, where B is the branching factor, and each iteration requires maximum one tree updating, which requires $f \cdot O(M)$. Overall, the proposed method presents a computational complexity which is bounded by $O(|X|M^2)$.

In practice, using C# on a 8GB Ram Laptop, Intel i7-2640M CPU @ 2.8GHz, loading all features takes approximately 25 seconds while all other steps can run in real-time.

5 CONCLUSIONS

In this work we proposed a flexible framework for retrieving diverse social images of landmarks by exploiting an outlier prefiltering process and hierarchical clustering using textual, visual and user credibility information. Moreover, we proposed to apply the concept of Relevance Feedback in a novel way for diversification, by showing that both diversity and relevance of the retrieved images can be further refined by exploiting users' judgments on the results produced by the algorithm. Thanks to its flexibility we show not only that the proposed approach is able to include the RF paradigm in a clustering approach and improve the obtained results, but also that the proposed RF setup can reduce the number of the required iterations with the user, thanks to the introduction of an additional type of feedback.

Experimental results performed on the MediaEval 2015 "Retrieving Diverse Social Images" dataset show that the proposed framework can achieve very good performance in both cases of automatic

diversification or by exploiting the novel RF paradigm, improving state-of-art performance. Future work will be devoted to extend the proposed method allowing the retrieval of diverse images on different contexts, such as social events.

REFERENCES

[1] M. R. Anderberg. 1973. *Cluster Analysis for Applications.* Academic Press.

[2] J. Bian, Y. Yang, H. Zhang, and T. S. Chua. 2015. Multimedia Summarization for Social Events in Microblog Stream. *IEEE Transactions on Multimedia* 17, 2 (Feb 2015), 216–228. https://doi.org/10.1109/TMM.2014.2384912

[3] G. Boato, D.-T Dang-Nguyen, O. Muratov, N. Alajlan, and F. G. B. De Natale. 2015. Exploiting visual saliency for increasing diversity of image retrieval results. *Multimedia Tools and Applications* (2015), 1–22.

[4] B. Boteanu, I. Mironica, and B. Ionescu. 2014. A Relevance Feedback Perspective to Image Search Result Diversification. In *IEEE International Conference on Computer Vision*. 47–54.

[5] B. Boteanu, I. Mironica, and B. Ionescu. 2015. Hierarchical clustering pseudo-relevance feedback for social image search result diversification. In *IEEE International Workshop on Content-Based Multimedia Indexing*. 1–6. https://doi.org/10.1109/CBMI.2015.7153613

[6] J. Carbonell and J. Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *ACM SIGIR Conference on Research and Development in Information Retrieval*. 335–336.

[7] T. Chen, K.-H. Yap, and D. Zhang. 2014. Discriminative Soft Bag-of-Visual Phrase for Mobile Landmark Recognition. *IEEE Transactions on Multimedia* 16, 3 (2014), 612–622.

[8] Y. Chen, X. S. Zhou, and T. S. Huang. 2001. One-class SVM for learning in image retrieval. In *IEEE International Conference on Image Processing*, Vol. 1. 34–37.

[9] N. Dalal and B. Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 886–893.

[10] D.-T. Dang-Nguyen, G. Boato, F.G.B. De Natale, L. Piras, G. Giacinto, F. Tuveri, and M. Angioni. 2015. Multimodal-based Diversified Summarization in Social Image Retrieval. In *MediaEval*, Vol. 1436.

[11] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F.G.B. De Natale. 2015. A Hybrid Approach for Retrieving Diverse Social Images of Landmarks. In *IEEE International Conference on Multimedia and Expo*.

[12] V. de Weijer, C. Schmid, J. Verbeek, and D. Larlus. 2009. Learning Color Names for Real-world Applications. *IEEE Transactions on Image Processing* 18, 7 (2009), 1512–1523.

[13] A. L. Gînscă, A. Popescu, B. Ionescu, A. Armagan, and I. Kanellos. 2014. Toward an Estimation of User Tagging Credibility for Social Image Retrieval. In *ACM International Conference on Multimedia*. 1021–1024.

[14] D. Giordano, S. Palazzo, and C. Spampinato. 2016. A diversity-based search approach to support annotation of a large fish image dataset. *Multimedia Systems* 22, 6 (01 Nov 2016), 725–736.

[15] J.-T. Huang, C.-H. Shen, S.-M. Phoong, and H. Chen. 2005. Robust measure of image focus in the wavelet domain. In *Intelligent Signal Processing and Communication Systems*. 157–160.

[16] Z. Huang, B. Hu, H. Cheng, H. T. Shen, H. Liu, and X. Zhou. 2010. Mining Near-duplicate Graph for Cluster-based Reranking of Web Video Search Results. *ACM Transactions on Information Systems* 28, 4 (2010), 22:1–22:27.

[17] B. Ionescu, A.-L. Gînscă, B. Boteanu, A. Popescu, M. Lupu, and H. Müller. 2015. Retrieving Diverse Social Images at MediaEval 2015: Challenge, Dataset and Evaluation. In *MediaEval*, Vol. 1436.

[18] B. Ionescu, A. Popescu, M. Lupu, A. L. Gînscă, and Müller. 2014. Retrieving Diverse Social Images at MediaEval 2014: Challenge, Dataset and Evaluation. In *MediaEval*.

[19] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei. 2015. Author Topic Model-Based Collaborative Filtering for Personalized POI Recommendations. *IEEE Transactions on Multimedia* 17, 6 (June 2015), 907–918. https://doi.org/10.1109/TMM.2015.2417506

[20] L. S. Kennedy and M. Naaman. 2008. Generating Diverse and Representative Image Search Results for Landmarks. In *ACM International Conference on World Wide Web*. 297–306.

[21] D.-H. Kim, C.-W. Chung, and K. Barnard. 2005. Relevance feedback using adaptive clustering for image similarity retrieval. *Journal of Systems and Software* 78, 1 (2005), 9 – 23. https://doi.org/10.1016/j.jss.2005.02.005

[22] J. Laaksonen, M. Koskela, and E. Oja. 2002. PicSOM-self-organizing image retrieval with MPEG-7 content descriptors. *IEEE Transactions on Neural Networks* 13, 4 (2002), 841–853.

[23] S. Lazebnik, C. Schmid, and J. Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2169–2178.

[24] S. Liang and Z. Sun. 2008. Sketch retrieval and relevance feedback with biased SVM classification. *Pattern Recognition Letters* 29, 12 (2008), 1733 – 1741. https://doi.org/10.1016/j.patrec.2008.05.004

[25] D. Lu, X. Liu, and X. Qian. 2016. Tag-Based Image Search by Social Re-ranking. *IEEE Transactions on Multimedia* 18, 8 (Aug 2016), 1628–1639. https://doi.org/10.1109/TMM.2016.2568099

[26] Z. Lu and H.H.S. Ip. 2010. Combining Context, Consistency, and Diversity Cues for Interactive Image Categorization. *IEEE Transactions on Multimedia* 12, 3 (2010), 194–203.
[27] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada. 2001. Color and Texture Descriptors. *IEEE Transactions on Circuits and Systems for Video Technology* 11, 6 (2001), 703–715.
[28] I. Mironica, B. Ionescu, and C. Vertan. 2012. Hierarchical clustering relevance feedback for content-based image retrieval. In *IEEE International Workshop on Content-Based Multimedia Indexing*. 1–6.
[29] T. Ojala, M. Pietikinen, and D. Harwood. 1994. Performance Evaluation of Texture Measures with Classification based on Kullback Discrimination of Distributions. In *IAPR International Conference on Pattern Recognition*. 582–585.
[30] M. Paramita, M. Sanderson, and P. Clough. 2009. Diversity in Photo Retrieval: Overview of the ImageCLEF Photo Task 2009. In *International Conference on Cross-language Evaluation Forum: Multimedia Experiments*.
[31] L. Piras and G. Giacinto. 2009. Neighborhood-based feature weighting for relevance feedback in content-based retrieval. In *IEEE International Workshop on Image Analysis for Multimedia Interactive Services*. 238–241.
[32] L. Piras and G. Giacinto. 2017. Information fusion in content based image retrieval: A comprehensive overview. *Information Fusion* 37 (2017), 50 – 60. https://doi.org/10.1016/j.inffus.2017.01.003
[33] X. Qian, X. Tan, Y. Zhang, R. Hong, and M. Wang. 2016. Enhancing Sketch-Based Image Retrieval by Re-Ranking and Relevance Feedback. *IEEE Transactions on Image Processing* 25, 1 (Jan 2016), 195–208. https://doi.org/10.1109/TIP.2015.2497145
[34] X. Qian, Y. Xue, X. Yang, Y.Y. Tang, X. Hou, and T. Mei. 2015. Landmark Summarization With Diverse Viewpoints. *IEEE Transactions on Circuits and Systems for Video Technology* 25, 11 (2015), 1857–1869. https://doi.org/10.1109/TCSVT.2014.2369731
[35] S.S. Ravindranath, M. Gygli, and L. van Gool. In *MediaEval*.
[36] S. Rudinac, A. Hanjalic, and M. Larson. 2013. Generating Visual Summaries of Geographic Areas Using Community-Contributed Images. *IEEE Transactions on Multimedia* 15, 4 (2013), 921–932.
[37] Y. Rui, T. S. Huang, and S. Mehrotra. 1997. Content-Based Image Retrieval with Relevance Feedback in MARS. In *IEEE International Conference on Image Processing*. 815–818.
[38] Y. Rui, T. S. Huang, and S. Mehrotra. 1998. Relevance feedback: A power tool in interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 8, 5 (September 1998), 644–655.
[39] S. Sabetghadam, J.R.M. Palotti, N. Rekabsaz, M. Lupu, and A. Hanbury. 2015. TUW @ MediaEval 2015 Retrieving Diverse Social Images Task. In *MediaEval*, Vol. 1436.
[40] I. Simon, N. Snavely, and S. M. Seitz. 2007. Scene Summarization for Online Image Collections. In *IEEE International Conference on Computer Vision*. 1–8.
[41] B. Thomee and M. S. Lew. 2012. Interactive search in image retrieval: a survey. *International Journal of Multimedia Information Retrieval* 1, 1 (2012), 71–86.
[42] R. Tronci, G. Murgia, M. Pili, L. Piras, and G. Giacinto. 2013. ImageHunter: A Novel Tool for Relevance Feedback in Content Based Image Retrieval. In *New Challenges in Distributed Information Filtering and Retrieval*. Vol. 439. 53–70.
[43] C.-M. Tsai, A. Qamra, E.Y. Chang, and Y.-F. Wang. 2006. Extent: Interring image metadata from context and content. In *IEEE International Conference on Multimedia and Expo*. 1270–1273.
[44] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol. 2009. Visual Diversification of Image Search Results. In *ACM International Conference on World Wide Web*. 341–350.
[45] T. Wang, Y. Rui, S.-M. Hu, and J.-G. Sun. 2003. Adaptive tree similarity learning for image retrieval. *Multimedia System* 9, 2 (2003), 131–143.
[46] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo.. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3485–3492.
[47] E.S. Xioufis, A. Popescu, S. Papadopoulos, and I. Kompatsiaris. USEMP: Finding Diverse Images at MediaEval 2015. In *MediaEval*.
[48] M. Zaharieva and L. Diem. 2015. MIS @ Retrieving Diverse Social Images Task 2015. In *MediaEval*, Vol. 1436.
[49] L. Zhang, F. Lin, and B. Zhang. 2001. Support vector machine learning for image retrieval. In *IEEE International Conference on Image Processing*, Vol. 2. 721–724.
[50] R. Zhang and Z. Zhang. 2005. FAST: Toward more effective and efficient image retrieval. *Multimedia System* 10, 6 (2005), 529–543.
[51] T. Zhang, R. Ramakrishnan, and M. Livny. 1996. BIRCH: An Efficient Data Clustering Method for Very Large Databases. In *ACM SIGMOD International Conference on Management of Data*. 103–114.
[52] L. Zhu, J. Shen, H. Jin, L. Xie, and R. Zheng. 2015. Landmark Classification With Hierarchical Multi-Modal Exemplar Feature. *IEEE Transactions on Multimedia* 17, 7 (2015), 981–993.

Received November 2016; revised April 2017; accepted May 2017