A HYBRID APPROACH FOR RETRIEVING DIVERSE SOCIAL IMAGES OF LANDMARKS

Duc-Tien Dang-Nguyen[1], Luca Piras[1], Giorgio Giacinto[1], Giulia Boato[2], Francesco G. B. De Natale[2]

[1] DIEE - University of Cagliari, Italy
{ductien.dangnguyen, luca.piras, giacinto}@diee.unica.it
[2] DISI - University of Trento, Italy
boato@disi.unitn.it, denatale@ing.unitn.it

ABSTRACT

In this paper, we present a novel method that can produce a visual description of a landmark by choosing the most diverse pictures that best describe all the details of the queried location from community-contributed datasets. The main idea of this method is to filter out non-relevant images at a first stage and then cluster the images according to textual descriptors first, and then to visual descriptors. The extraction of images from different clusters according to a measure of user's credibility, allows obtaining a reliable set of diverse and relevant images. Experimental results performed on the MediaEval 2014 "Retrieving Diverse Social Images" dataset show that the proposed approach can achieve very good performance outperforming state-of-art techniques.

Index Terms— Social Image Retrieval, Diversity

1. INTRODUCTION

Ten years after the rise of social networks and photo storage services such as Facebook and Flickr, the number of online pictures has incredibly increased, approaching in year 2014 one trillion uploads. Thus, the need for efficient and effective photo retrieval systems has become crucial. However, current photos search engines, e.g., Bing or Flickr, mainly provide users with exact results for the queries, which are basically the visually or verbally best matches and usually provide redundant information.

Diversity has been demonstrated to be a very important aspect of results expected by users, to achieve a comprehensive and complete view on the query [1]. 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 of image search results has been studied by many researchers, and some international challenges have been also proposed around the problem (ImageCLEF [2] and MediaEval Retrieving Diverse Social Images Task [3]). In particular, the MediaEval task focuses on social photo retrieval and in particular aims at providing a more complete visual description of a given location.

(a) Results from Flickr. (b) Diversified results.

Fig. 1. An example of the first 10 results retrieved by Flickr default search for query Rialto and the first 10 resutls of the proposed approach where the landmark is represented in a more diversified and complete way.

In this context, the results are supposed not only to provide relevant images of one specific landmark but also complementary views of it with different perspectives, various day-times (e.g., night and day), which may provide comprehensive understanding of the queried location (see Figure 1).

In this paper we address this problem, aiming at producing a visual description of the landmark by choosing the most diverse pictures that best describe all its details (e.g., in the case of a building: different external views, details, interior pictures, etc.). We propose a hybrid framework that exploits textual, visual and user credibility information of social images. Starting from a set of images of a landmark, retrieved through tag information, the first step of the proposed method is to filter out the irrelevant pictures: photos taken in the queried location but that do not show the landmark in foreground (e.g., close-up pictures of people), and blurred or out of focus images are removed. As a second step, we use a hierarchical clustering algorithm that is designed to ensure diversity, that first performs clustering according to textual information, then refines the results by visual information. Finally, in the third step, we produce the summary for the queried location by selecting representative images from the clusters based on the user credibility information, which mainly represents how good the image-tag pairs uploaded from users are.

In order to evaluate the proposed framework, we run the experiments on the MediaEval 2014 "Retrieving Diverse Social Images" dataset [3]. A preliminary version of this approach has been presented during the competition [4].

The structure of the paper is the following: in Section 2 related work is briefly described; in Section 3 the proposed framework 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

Current works in this field have considered relevance and diversity as two core criteria of efficient landmark image retrieval systems. Relevance was commonly estimated based on textual information, e.g., from the photo tags [5], and many of current search engines are still mainly based on this information. However, textual information are normally inaccurate, e.g., users commonly tag the entire collection with only one tag. Some other works have exploited the improvement of low-level image descriptors, e.g. SIFT [6], or a fusion of textual and visual information to improve the relevance [7]. Low-level visual descriptors, however, often fail to provide high-level understanding of the scene. Thus, extra information is required, for example, in [8], the authors exploited GPS information to provide users with accurate results for their queries.

Diversity is usually improved by applying clustering algorithms which rely on textual or/and visual properties [7]. In [9], the authors define a criterion to measure the diversity of results in image retrieval and attempt to optimize directly this criterion. Some approaches have used a "canonical view" [10] based on unsupervised learning to diversify the search results. Some other exploited visual saliency to re-rank top results and improve diversification [11]. Recently, some methods have exploited the participation of humans by collecting the feedbacks of the results to improve the diversification [12].

In [13], a novel information has been introduced: user credibility, which mainly represents how good the image-tag pairs uploaded from users are. This is extracted from a large amount of annotated data and can be integrated with different cues to improve the landmark search performance, as done in the proposed approach.

MediaEval Benchmarking Initiative for Multimedia Evaluation organizes since 2013 a task on retrieving diverse social images [3], by publishing a large collection of landmark images with the ground truth annotated by experts. Assessment of the proposed method is performed on this dataset.

3. METHODOLOGY

The proposed method is made up of three main steps, as illustrated in Figure 2.

Starting from an initial set of images retrieved through tag information, the first step is to filter out non-relevant images which are taken outside of the queried location, or taken in that place but do not show the landmark in foreground (e.g., close-up pictures of people), or that are blurred or out of focus. Shown in panel (a) and (b) are the initial images retrieved from the query "Colosseum" using the Flickr default search, and the filtering results, respectively.

Next, we apply a hierarchical clustering algorithm that is designed to ensure the diversity. Clustering is applied on textual information to construct a clustering tree, and the tree is then refined based on visual information. Shown in panel (c) are the clusters after clustering step, where each cluster represents a different aspect of the queried landmark. Images in the same cluster are not only visually similar but also coherent in the textual information.

Finally, the summarization step is applied by sorting the clusters based on their size, then from each cluster the image with the highest visual score, representing user credibility, is selected. Illustrated in panel (d) is an example of the final result of the process where the queried location has been succesfully summarized. It can be noticed that the final set provides a diversified view of the landmark, with images which are relevant but represent various viewpoints (e.g., from inside, from outside), day and night pictures, details, etc.

The details of each step will be described in the following sections.

3.1. Filtering outliers

The goal of this step is to filter out outliers by removing images that can be considered as non-relevant. We consider an image as non-relevant by defining the following rules:

1. It contains people as 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, Luxand FaceSDK[1] is used as a face detector. A detected face is only confirmed as a human face after checking its color in H channel (in the HSV color space) in order to avoid mis-detection from 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 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 $R = 6356.752km$ 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 wavelet coefficients and compare it to a threshold, following [14].

[1] luxand.com

Fig. 2. Schema of the proposed framework

4. It received very few number of views. Since we are working on social images datasets (e.g., Flickr), if an image received a few 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 images left are considered as relevant and are passed to the next step.

3.2. Clustering by using BIRCH algorithm on textual-visual descriptors

In this step, we use the Balanced Iterative Reducing and Clustering (BIRCH) algorithm [15] on textual and visual descriptors to diversify the results by clustering the filtered images from the previous step.

In order to combine textual and visual information together, we build a tree based on textual information, then refined it based on visual information. The reason behind this is that as soon as the pictures are taken in the same place, their visual similarity is high and confusion may arise. Thus, by reducing the number of images to be visually processed into smaller more coherent subsets can make the visual processing problem less expensive and more likely to yield precise results. BIRCH can typically find a good clustering with a single scan of the dataset and further improve the quality with a few additional scans. Furthermore, it can also handle noise effectively. Thus, we decided to use BIRCH [15] for this step.

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 procedure is summarized in Algorithm 1. In phase 1, the CF tree is built using the textual feature vectors X. CF tree is a height-balanced tree which is based on two parameters: branching factor B and 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. If it is smaller than T, a test is performed to see whether the vector belongs to the candidate cluster or not. If not, a new cluster is created and added to the father node. Then, any node that contains more than B children is splitted.

BIRCH provides an optional step to "restructure" the tree obtained in the first step in order to obtain a more tidy and compact tree. We have done it by replacing the text features with visual ones: for each node, its center and radius are recomputed based on visual feature vectors V instead of the former textual feature vectors X. T is updated with the largest radius from leaf clusters. Phase 2 of the algorithm is then applied by rebuild the tree after increasing T and keeping the same B. Finally, the clusters are extracted from the CF tree by applying phase 3 and phase 4 of the algorithm.

After this step, images that are visually similar and have the same context (textual information) are grouped into the same cluster.

3.3. Summarization based on user credibility information

From the clusters obtained from the clustering step, representative images that best describe the queried location are selected. Here, we propose a novel way for choosing such images by exploiting the user credibility information.

Credibility scores are estimated by exploiting ImageNet, a manually labelled dataset of 11 million images of around 22000 concepts. For each user, at most 300 images which have tags that matched with at least one of the ImageNet con-

Algorithm 1 Image clustering according to BIRCH [15]

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 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 algorithm on CF leaves to form the set of clusters K.
Phase 4 Redistribute the data points to its closest seed to obtain a set of new clusters.

cepts are selected. Tags are then analyzed against the corresponding concepts to obtain individual relevance-score. Details of how the score was built can be seen in [13]. Here, we use the visual score of a user, which represents how relevant the images uploaded by that user are, to decide the representative images of each cluster.

To select the best images that can summarize 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 are found (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 are more than one image from that user, the image closest to the centroid is selected. If more than one image has to be extracted from each cluster, the second image is the one which has the largest distance to the first image, the third image is chosen as the image with the largest distance to both the first two images, and so on.

The distance between two images i and j is computed by fusing visual distance with the focus value (computed from the filtering step): $d_s(i,j) = \alpha \cdot ||\overrightarrow{v_i} - \overrightarrow{v_j}|| + \beta \cdot |f_i - f_j|$, where $\overrightarrow{v_i}$, $\overrightarrow{v_j}$ and f_i, f_j are the visual descriptors and focus values of images i and j, respectively, $||\cdot||$ is the Euclidean distance, $|\cdot|$ is the absolute value, α, β are the normalizing constants.

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 2014 "Retrieving Diverse Social Images" [3]. This dataset is built from around 45.000 images from 153 locations spread over 35 countries allover the world. The images were collected from Flickr by querying on the location names using Flickr default algorithm. Flickr metadata of each image (e.g., photo title, photo description, photo id, tags, etc.) are also provided together with the content descriptors which consist on visual, textual and user credibility information.

Table 1. Visual descriptors evaluated in the test. The postfix '3x3' in the name of some descriptors denotes that these descriptors are computed by concatenate the descriptors extracted from 3×3 non-overlapping blocks of the image. Descriptors in bold provide the best performance in terms of cluster recall.

Provided descriptors [3]	Extracted descriptors [16]
- Global Color Naming Histogram (CN)	- GIST
- CN 3x3	**- HOG2x2**
- Global Histogram of Oriented Gradients (HOG)	**- Dense SIFT**
	- Sparse SIFT histograms
- Global Color Moments (CM)	**- LBP with uniform patterns**
- CM 3x3	- SSIM: Self-similarity descriptors
- Global Color Structure Descriptor (CSD)	- Tiny Images
- Global Statistics on Gray Level Run Length Matrix (GLRLM)	- Line Features
	- Texton Histograms
- GLRLM 3x3	- Color Histograms
- Global Locally Binary Patterns (LBP)	- Geometric Probability Map
- LBP 3x3	- Geometry Specific Histograms

The images are annotated with respect to both relevance and diversity by experts with advanced knowledge of the locations. Relevant images are grouped into 20 to 25 clusters, where each cluster depicts an aspect of the queried location (e.g., side-view, close-up view, drawing, sketch, etc.).

The dataset was also splitted into developing set (devset) and testing set (testset), containing 30 and 123 locations, respectively. Devset is mainly used for training and testset is used for testing.

Performance with respect to relevance and diversity are assessed using the following standard metrics:

Precision. Relevance is measured by precision at N ($P@N$), defined as: $P@N = \frac{N_r}{N}$, where N_r is the number of relevant images from the first N ranked results.

Cluster recall. Diversity is assessed with cluster recall at N ($CR@N$), defined as: $CR@N = \frac{N_c}{N_{tc}}$, where N_c is the number of clusters found from 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: $F1@N = 2 \cdot \frac{P@N \cdot CR@N}{P@N + CR@N}$.

In the experiments, all of the above measures are considered with various cut off points $N = 5, 10, 20, 30, 40, 50$.

4.2. Experiments on different visual descriptors

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

As mentioned in Section 4.1, descriptors, together with the metadata, are provided with the images. However, these descriptors, which are listed in the left column in Table 1, mainly represent the global information of the image and do

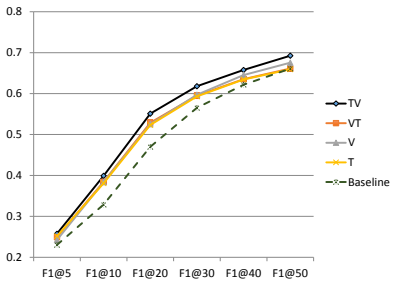

Fig. 3. Clustering step evaluation.

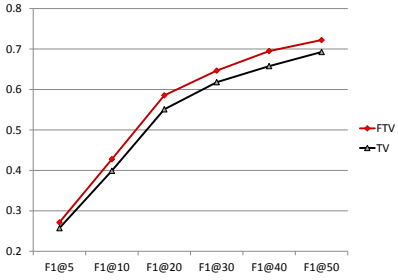

Fig. 4. Filtering step evaluation.

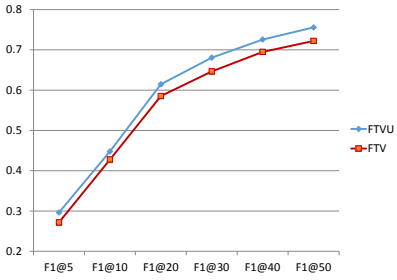

Fig. 5. Summarization step evaluation.

not take into account the information at a more detailed level. Thus, we also extracted and used other 12 visual descriptors, following [16], collecting in total 22 visual descriptors for our evaluation.

In order to find the best combination of both descriptors and parameters (number of clusters, inner parameters of BIRCH algorithm), we performed a test on all relevant images in the devset, and measured the accuracy based on cluster recall. Only visual information is used and the centroids of the clusters (after the clustering step) are selected as representative images.

Best performances in terms of cluster recall (for all cut off points $N = 5, 10, 20, 30, 40, 50$) were obtained by using the following visual descriptors: global color naming histogram, histogram of oriented gradients 2x2, dense SIFT, locally binary pattern with uniform patterns, and global color structure descriptor. Thus, for the rest of the experiments, we used these visual descriptors and the tuned parameters. Table 1 lists all 22 descriptors which have been tested, where the bold ones are the selected descriptors.

4.3. Evaluation on the proposed method

Clustering step evaluation

In order to evaluate the clustering step, we performed a test using four different configurations as follows: using only visual descriptors (denoted as V), using only text descriptors (denoted as T), clustering based on textual descriptors and then refined based on visual descriptors (denoted as TV), clustering based on visual descriptors and then refined based on textual descriptors (denoted as VT). In this test, filtering step was not applied and the centroids of the clusters were selected as representative images, i.e., without using 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 3, where it can be easily seen that at all cut off points, the TV configuration outperforms the others, supporting the considerations done in Section 3.2 on using textual information prior to visual information.

Filtering step evaluation

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%** accuracy, when **41%** of the non-relevant images were detected using these thresholds: (i) the face size is bigger than 10% with respect to the size of the image, (ii) images that were shot farther than 15kms, (iii) images that have less than 20 views, and (iv) images that have the f-focus value (at the first stage) smaller than 20.

Table 2. Results on different cut off points.

	P@10	P@20	P@30	CR@10	CR@20	CR@30	F1@10	F1@20	F1@30
FTVU	0.873	0.858	0.822	0.301	0.479	0.581	0.448	0.615	0.681
FVTU	0.859	0.849	0.819	0.296	0.465	0.553	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	0.703	0.260	0.424	0.533	0.384	0.525	0.597
T	0.727	0.718	0.714	0.262	0.419	0.519	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

Removing non-relevant images at the early stage significantly improves the performance of the final set of images. Shown in Figure 4 are the $F1@N$ of the proposed method with and without applying the filtering step (denoted as FTV and TV, respectively). Configuration TV, explained in the previous test, was used.

Summarization step evaluation

The importance of user credibility information was assessed by running the best configuration of previous tests with and without using the visual score information (FTVU and FTV, respectively). Shown in Figure 5 are the $F1@N$ at different cut off points, showing that using the user credibility information the proposed method can summarize the queried location better.

Detailed results of all tested configurations on cut off points $N = 10, 20, 30$ are reported in Table 2, showing that configuration FTVU provides the best performance at all metrics.

4.4. Comparison with state-of-art methods

In this last experiment, we compared the proposed method with the three methods that achieved best performance in the

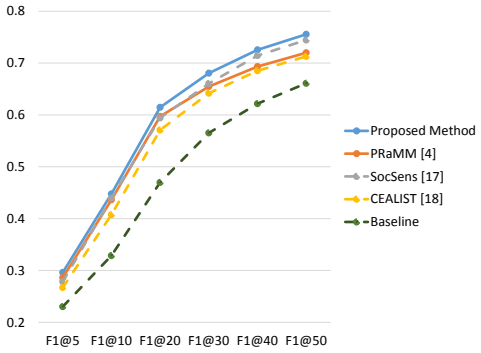

Fig. 6. Results on $F@N$.

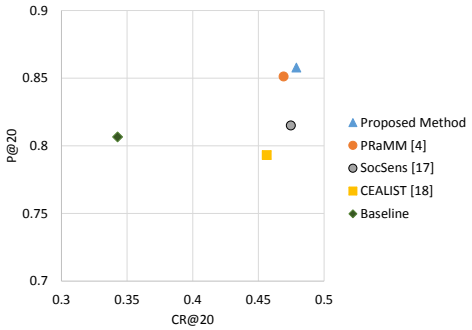

Fig. 7. Results on $P@20$ and $CR@20$.

MediaEval 2014 "Retrieving Diverse Social Smages" competition, namely PRa-MM [4], SocSens [17], and CEALIST [18]. Following the rules of the competition, we tuned the parameters and configurations using the images in the devset, and then applied the method to the testset. One of the compared methods was our preliminary study submitted to MediaEval 2014. In this study, the parameters and configurations were not optimized. The selection of visual descriptors was not implemented and the distance between two images in the summarization was only computed from visual descriptors.

Shown in Figure 6 is the $F1@N$ on all the cut off points, showing that the proposed method outperforms all other methods at all cut off points. Considering the official ranking metric of the competition, which is measured at the cut off point $N = 20$, the proposed method provides better performances on both $P@20$ and $CR@20$ (shown in Figure 7).

5. CONCLUSIONS

We proposed a novel approach for retrieving diverse social images of landmarks by exploiting an outlier prefiltering process and hierarchical clustering using textual, visual and user credibility information. Experimental results, performed on the MediaEval 2014 "Retrieving Diverse Social Images" dataset, show that the proposed method achieves very good precision and cluster recall, improving state-of-art performance. Future work will be devoted to extend the method allowing retrieval of diverse images also on different context, like social events.

Acknowledgement

This work is 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).

6. REFERENCES

[1] J. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," in *ACM SIGIR*, 1998, pp. 335–336.

[2] M. Paramita et al., "Diversity in photo retrieval: overview of the imageclef-photo task 2009," in *ImageCLEF*, 2009.

[3] B. Ionescu el al., "Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation," in *MediaEval 2014 Workshop*, Barcelona, Spain, 2014.

[4] D.-T. Dang-Nguyen et al., "Retrieval of diverse images by prefiltering and hierarchical clustering," in *MediaEval*, 2014, First place.

[5] C.-M. Tsai et al., "Extent: Inferring image metadata from context and content," in *ICME*, 2006, pp. 1270–1273.

[6] S. Rudinac et al., "Generating visual summaries of geographic areas using community-contributed images," *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 921–932, 2013.

[7] R. H. van Leuken et al., "Visual diversification of image search results," in *ACM WWW*, 2009, pp. 341–350.

[8] L. S. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *ACM WWW*, 2008, pp. 297–306.

[9] Z. Huang et al., "Mining near-duplicate graph for cluster-based reranking of web video search results," *ACM Transactions on Information Systems*, vol. 28, no. 4, pp. 22:1–22:27, 2010.

[10] I. Simon et al., "Scene summarization for online image collections," in *ICCV*, 2007, pp. 1–8.

[11] G. Boato et al., "Exploiting visual saliency for increasing diversity of image retrieval results," *ACM MTAP*, pp. 1–22, 2015.

[12] B. Boteanu et al., "A relevance feedback perspective to image search result diversification," in *ICCV*, 2014, pp. 47–54.

[13] A. L. Gînscă et al., "Toward Estimating User Tagging Credibility for Social Image Retrieval," in *ACM Multimedia*, 2014, pp. 1021–1024.

[14] J.-T. Huang et al., "Robust measure of image focus in the wavelet domain," in *ISPACS*, 2005, pp. 157–160.

[15] T. Zhang et al., "BIRCH: An Efficient Data Clustering Method for Very Large Databases," in *ACM SIGMOD*, 1996, pp. 103–114.

[16] J. Xiao et al., "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR*, 2010, pp. 3485–3492.

[17] E. Spyromitros-Xioufis et al., "Socialsensor: Finding diverse images at mediaeval 2014," in *MediaEval*, 2014.

[18] A. L. Gînscă et al., "Cea list's participation at the mediaeval 2014 retrieving diverse social images task," in *MediaEval*, 2014.