Multimed Tools Appl (2016) 75:5581–5602
DOI 10.1007/s11042-015-2526-4

Exploiting visual saliency for increasing diversity of image retrieval results

Giulia Boato · Duc-Tien Dang-Nguyen · Oleg Muratov · Naif Alajlan · Francesco G. B. De Natale

Received: 25 June 2014 / Revised: 20 January 2015 / Accepted: 19 February 2015 /
Published online: 6 March 2015

Abstract Diversification of search results allows for better and faster search, gaining knowledge about different perspectives and viewpoints on retrieved information sources. Recently various methods for diversification of image retrieval results have been proposed, mainly using textual information or techniques imported from the natural language processing domain. However, images contain much more information than their textual descriptions and the use of visual features deserves special attention in this context. Visual saliency provides information about parts of the image perceived as most important, which are instinctively targeted by humans when shooting a photo or looking at a picture. For this reason we propose to exploit such information to improve diversification of search results. To this purpose, we introduce a saliency-based method to re-rank the results of a query and we show that it can achieve significantly better performances as compared to the baseline approach. Experimental validation conducted on a number of queries applied to various datasets demonstrates the potential of the use of saliency information for the diversification of image retrieval results.

Keywords Visual saliency · Content-based image retrieval · Diversity

1 Introduction

Diversity of contents is an important feature and an added value in the Web, and in general in all applications characterized by a large amount of information coming from different sources [1]. It is the result of the large variety of situations, contexts, cultural backgrounds, religious and political beliefs, ideologies and time. Thus, to fully exploit the huge and

G. Boato (✉) · D.-T. Dang-Nguyen · O. Muratov · F. G. B. De Natale
Department of Information Engineering and Computer Science, University of Trento,
via Sommarive 9, Trento, Italy
e-mail: boato@disi.unitn.it

N. Alajlan
College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

ever increasing amount of information available on the Web, diversity has to be appropriately taken into account as a key instrument to achieve deeper understanding and reliable interpretation of the information and knowledge available. In the specific domain of media search, diversity is usually associated to a problem of perceptual diversification. Since search engines on the Web mainly exploit textual tags associated to images, they typically fail to provide this feature, thus retrieving many near duplicates. Instead, users would benefit of a higher variety of relevant results [15], especially when the query is poorly specified or ambiguous [26].

Diversification of results in media search engines is a relatively new area of research [24]. The importance of this research topic has been witnessed by the large participation of the research community to international challenges proposed around the image diversification problem by ImageCLEF 2009[1] and more recently by a new task within the MediaEval benchmarking framework.[2] Several techniques have been proposed to achieve this goal, mainly using textual information or algorithms imported from the natural language processing domain. Although image annotation could be an important source of information, quite often it turns out to be unreliable. For instance, user generated contents are often unannotated or sparsely annotated, thus making text-based approaches hardly applicable. Additionally, annotations may contain noisy or irrelevant data that in turn could produce irrelevant outputs. As a consequence, the degree of results diversification depends on how annotations grasp the content of the image both from visual and semantic points of view. On the other hand, images contain much more information than their textual descriptions and the use of visual features deserves special attention in this context. This is also proved by very recent results [3] where visual features have been proven to be fundamental to achieve a satisfactory image diversification. In terms of image search, a simple yet effective way to increase diversity is to ensure that duplicates or near-duplicates in the retrieved set are hidden from the user [30]. This approach however works as a posteriori filter on the result, while a mechanism to enforce diversification in the retrieval process would have more impact. An insight on the most significant approaches proposed so far will be presented in Section 2.

When dealing with visual perception of media objects, the concept of saliency is of paramount importance. Visual saliency provides information about the areas of an image perceived as most important and instinctively targeted by humans when shooting a photo or looking at a picture [12, 20, 21]. Intuitively, saliency can play an important role in the framework of diversification, by providing information on what the user perceives as the key subject of an image [22, 25], thus making it possible to focus the diversification on the most relevant contents. Stated another way, visual saliency can be considered as an additional dimension of the data implicitly embedded in a picture by its creator, which can be exploited for defining a higher dimensional feature space that allows more accurate description of images, emphasizing both semantic and visual diversity. What usually happens in content-based retrieval methods is that the content of the whole image is described in a uniform way. This approach ignores the obvious consideration that not all parts of an image have the same impact from the perceptual viewpoint. Typically an image represents a subject, which is probably the purpose for taking the picture, and a background, which represents the context in which the subject has been taken and can be more or less significant with respect to the subject. Let us consider the case of the query "Paris": pure visual diversity

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

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

applied to the whole image may result in different views of the same subject (for instance, the Eiffel tower) that differ on the dominant background area, while, focusing on the subject area, one can retrieve a higher variety of landmarks, been less biased by the differences in the background [11, 23]. Given the attitude of human beings to take photos according to specific, although often implicit, rules, it is commonly accepted that visual saliency provides a good approximation of what is intended to be the main subject of a picture.

In this paper we study the usefulness of visual saliency to increase diversity in image retrieval. We propose a method to re-rank the results of a query based on visual content, in order to achieve better diversity in top results. Then, we show how the introduction of a saliency-based modification of the re-ranking strategy can achieve significantly better performance as compared to the baseline approach. We will demonstrate that this allows achieving better diversification of the main subject of the picture (e.g., different viewpoints, different models of the same object, etc.), or vice versa providing different views of similar objects in different contexts (e.g., different backgrounds).

The rest of the paper is organized as follows: Section 2 provides an overview on recent related works; Section 3 describes the proposed approach detailing the saliency detection tool used, the representation of retrieved images, and the saliency-based re-ranking process; the evaluation of the proposed approach is illustrated in Section 4 followed by the conclusions in Section 5, where we provide some final considerations on the proposed idea and we highlight some future perspectives.

2 Related works

The idea of diversification of image retrieval results has been studied recently by many researchers [2, 16, 23] and the importance of this research topic is demonstrated also by the international challenges that are proposed around the problem. In 2009, ImageCLEF introduced the concept of diversity in the photo retrieval task, aiming at reducing duplicate images in search results [19]. More recently a new benchmarking task within MediaEval has been proposed around this topic [10]. The Retrieving Diverse Social Images Task focuses on the problem of result diversification in social photo retrieval and in particular aims at providing a more complete visual description of locations. Given a ranked list of location photos retrieved using text information, the participating systems were expected to provide a set of images that are at the same time relevant and maximally diverse (e.g., depict different views of the location at different times, from various perspectives, etc.).

A through comparison of different methods submitted to the ImageCLEF retrieval contest can be found in [27], including text-only, hybrid and pure content-based methods and showing that with current technologies hybrid approaches outperform text-only and content-based methods. A notable example of a hybrid method was presented by [4]. In their approach, unlike the many methods performing diversification as a post-processing step, the authors proposed a dynamic-programming-like ranking that jointly optimizes relevance and diversity measures. To this purpose, they use a broad variety of input features that include colour histograms, texture descriptors, bag of visual words, and text data. Another approach with similar characteristics can be found in [28]: here, unlike the above mentioned work, the authors used visual and textual features separately. Text features are responsible for the relevance by estimating the distance of tags, while visual features are used for diversification by maximizing the distance among candidate images. A pure text-based method was proposed by [32]. The authors presented a probabilistic model of image tags, with respect to the query that models both relevance and diversity. Increasing diversity without relevance deterioration is indeed a major issue [19]. A nice example focused on landmarks can be found in [11]. The main disadvantages of the above methods is that they rely on the semantic relationship of textual annotation, thus making them hardly applicable to unreliably annotated data.

An interesting approach dealing with unannotated data has been presented by [26]. The authors addressed the problem of diversification through automatic annotation of images based on their visual features. Text information is then used for creating a topic graph of the set. Finally, the results are diversified using a topic richness score, so that images with higher score appear at the top of the ranking. In addition, a topic coverage score is proposed, which measures the diversity of the image set and is based on the number of text-topics present in the results. Although this method is independent of image annotation, its performance is highly dependent on the results of the annotation prediction method.

The use of clustering techniques as a post-retrieval processing step for topic coverage enhancement has been proposed in the work by [13]. The authors performed comparison among several clustering strategies and analysed their effect on relevance and topic coverage. They also proposed a dynamic feature weighting technique that allows better fusion of features. Clustering is performed using a visual similarity measure based on low-level features and descriptors. Like in most content-based methods, all the content is treated uniformly, without differentiating between important areas and background.

Also the very recent Retrieving Diverse Social Images Task 2014 showed interesting results. A novel information was used in addition to visual and textual features: user credibility. Best results have been achieved by [3], by exploiting pre-filtering to reduce not-relevant images and then constructing a hierarchical tree allowing to cluster images with several criteria on visual and textual features.

Although saliency detection is still rarely used in multimedia applications ([29] recently proposed its application to photo collage generation), a first idea of using saliency information to re-rank retrieval results can be found in [9], where however the purpose was to improve the relatedness of top images and not their diversity. Indeed, the main goal was to improve visual consistency and attractiveness of top results, thus allowing also near-duplicates, which can be considered a concurrent objective with respect to diversification. Moreover, the authors are concerned with salient images in a group of pictures, and not with salient regions in each individual image.

3 Proposed approach

In this work we propose to exploit visual saliency information to improve diversification of the content in search results. This section will provide a detailed description of the proposed method, which is composed by three steps. We start with the description of the visual saliency map extraction method (Section 3.1). Then, we describe the features used for diversification and how they are related to the relevant application (Section 3.2). Finally, we present the proposed saliency-based re-ranking approach (Section 3.3).

3.1 Visual saliency detection

The overall purpose of the proposed method is to provide an innovative way to exploit saliency information to diversify image search results. Therefore, the first step to be performed is to extract a meaningful saliency map, i.e., to partition the image in areas characterized by higher or lower perceptual relevance. To achieve this goal, we implemented a

Table 1 EDISON segmentation parameters used in this work

Minimum region area	imheight · imwidth · 0.005
Spatial bandwidth	10
Range bandwidth	7.5
Gradient window radius	2
Mixture parameter	0.3
Edge strength threshold	0.7

bottom-up saliency detection method based on our previous work [18], introducing some modification to make it more suitable for the current application. This method is based on the analysis of low-level features extracted from a segmented image. The application of saliency in the segmented domain instead of in the pixel domain allows performing the classification at a higher level, thus achieving a higher consistency of the generated map with the objects present in the scene. Furthermore, it reduces the amount of noise in the estimation, due to the "averaging" effect within segments. In this work, we use the EDISON tool for segmentation[3] because of public availability of source code and satisfactory performance in terms of both accuracy and computational load. The tool is based on mean-shift segmentation, and requires the setting of various parameters. To achieve better object shape estimation the default parameters were tuned as reported in Table 1.

A schema of the extraction process is depicted in Fig. 1. In this model saliency is mostly derived from visual features described below. Some of these features cannot be extracted directly from segment data. Their values are computed first on the whole image, and segment-wise level is then obtained by averaging the feature value over that segment. Since visual saliency is commonly associated to the presence of irregularities, the selected features aim at detecting such characteristics.

Features extracted from the entire image include luminance contrast and center-surround histograms. Human attention is sensitive to contrast, thus luminance contrast is included into the proposed model and measured on a downscaled version of the image (by factor 8). The motivation is that maximum contrast values are usually observed on edges and glare spots, while downscaling allows to catch global contrast changes. The luminance contrast LC is computed as follows:

$$LC(x, y) = \sum_{m} \sum_{n} \frac{|L(x, y) - L(x+m, y+n)|}{\sqrt{m^2 + n^2}} \tag{1}$$

where $L(x, y)$ is the luminance value of the pixel with coordinates (x, y), and $m, n \in \{-2, -1\} \cup \{1, 2\}$ denote the relative coordinates of neighboring pixels.

Center-surround histograms allow to measure the distance between foreground and background. The underlying idea is that usually the histogram of the foreground object has a larger extent than its surroundings. Following [17], the input image is scanned by two rectangular windows R_f and R_s, where R_f is a notch inside the window R_s. The distance of foreground and surrounding histograms is computed as follows:

$$dist\left(R_s, R_f\right) = \frac{1}{2} \sum \frac{\left(R_f^i - R_s^i\right)^2}{R_f^i + R_s^i} \tag{2}$$

[3] http://coewww.rutgers.edu/riul/research/code/EDISON/

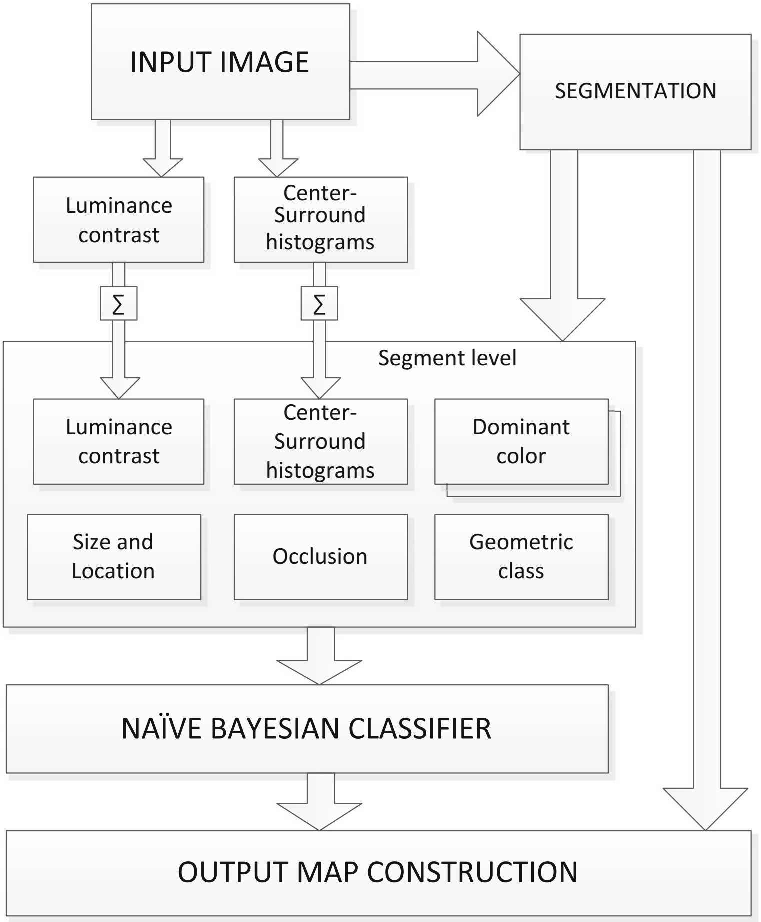


Fig. 1 Saliency detection flow: several features are computed on the whole image and on the segmented one in order to collect all information required to define an effective saliency map. Saliency map extraction is the first step of the proposed method and is needed for the subsequent visual features extraction on salient and not salient parts of the image

where R_s^i, R_f^i are surrounding and foreground histograms, respectively. Histogram distances are computed at each scale {0.3, 0.7} and aspect ratio {0.5, 1, 1.5}. Finally, they are normalized and summed into a single map. An average value of each global feature is assigned to each segment of the input image.

Local features, computed on segments, include the following descriptors: dominant color, spatial location, size, geometric class and occlusion. Colors have a great impact on the perception of objects. The dominant color is computed in a 12 tone color space following [6], where it is proved that some colors are more likely to attract attention than others. This feature is used in two ways: i) in conjunction with psychological studies on the impact of certain colors for human-attention and associating a value proportional to the color saliency defined in [6], and ii) for detecting segments with colors different from others after normalization over all segments. In particular, for each color tone a normalized weight w_c is computed as follows:

$$w_c = \frac{w'_c - min\left(w'_i\right)}{max\left(w'_i\right) - min\left(w'_i\right)} \tag{3}$$

where w'_c is the weight before normalization and it is defined as follows:

$$w'_c = \frac{1}{\sum_{i \in S_c} a_i} \tag{4}$$

where S_c is a set of all segments assigned to color c, a_i is the area of segment S_i.

The spatial location is based on the observation that amateurs tend to place the most interesting part of the image into the center. It is computed as the location of segment's center of mass. The location M_i of the segment S_i is computed as follows:

$$M_i = \left\{ \left[\sum_{(x,y) \in S_i} \left(\frac{mx}{2} - x \right) \right]^2 + \left[\sum_{(x,y) \in S_i} \left(\frac{my}{2} - y \right) \right]^2 \right\}^{\frac{1}{2}} \tag{5}$$

where mx and my are the image dimensions. As far as the size is concerned, the object of interest usually occupies a significant portion of the image. Thus it is unlikely that a very small segment is salient. The size plays the role of a filter, avoiding that larger regions (likely belonging to the background) or smaller ones (likely caused by noise or irrelevant details) are classified as salient.

In order to improve the accuracy of the map extraction with respect to [18] we introduced also the detection of the geometric class of each segment. To this purpose, we make use of the classifier proposed by [8], which provides classes such as ground, sky, diagonal and vertical planes. This feature allows the discrimination between planes that usually belong to background (e.g., sky, ground), and planes that belong to background or foreground with equal probability. We finally introduce an occlusion feature, to detect whether a segment is occluded by another. The intuition here is that usually, the main object of a scene is placed in a frontal plane. Thus, whenever a large segment is occluded by several smaller segments, it is unlikely to be the foreground. This is achieved by comparing the locations of the center of mass of neighboring regions, their size and shapes. Firstly, we compute the spread of each segment approximating with a rectangle the occupied area. Then, if two segments have overlapping regions, occlusion is detected by thresholding the area of their intersection.

All features described above allow representing each segment with a 12 element vector. To perform the binary classification of salient vs. non-salient segments, a Naive Bayesian classifier is used. First, we perform a training over a dataset of more than 700 images associated to ground-truth data. The input of the classifier is the vector of features assigned to each segment, while the output is the probability of the segment to be salient. The final saliency map is constructed by combining such probabilities with the segmentation map. Since our re-ranking method needs a hard classification, the resulting probabilistic map is finally binarized using a simple thresholding with a value 0.5 (see some examples in Fig. 2).

3.2 Visual features

In order to quantitatively measure the visual dissimilarity among images, it is necessary to define a set of features that efficiently encode the perceptual appearance of visual data. In this work we rely on low level features that are correlated with human vision system (HVS)

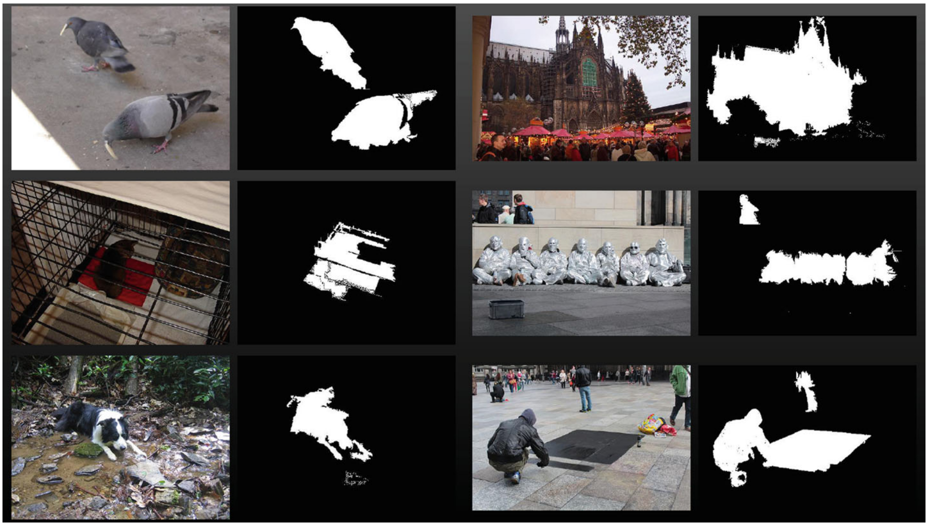


Fig. 2 Examples of saliency maps extracted from a set of images and thresholded with a value 0.5 in order to differentiate between foreground (salient area) and background (not salient areas)

characteristics. In particular, each image is described by 6 features, namely: foreground and background color histograms, foreground and background orientation histograms, foreground size and foreground location. In the following we provide further detail about this description while the overall scheme is shown in Fig. 3. We define as foreground and background the salient and not salient areas, respectively, detected by segment-based saliency extraction method described in Section 3.1.

Colors are recognized to be one of the most important perceptual features of images. In particular, color histograms provide a meaningful and convenient representation, accounting for relatively fast processing and easy comparison. Furthermore, the use of color histograms in previous works demonstrated their good applicability for the task of diversification. Color histograms can be applied to different color spaces and with different chromatic resolutions. Some works propose the use of entire full-color RGB color space, while others use alternative color representations such as L*a*b* or HSV, with different numbers of bins. In this work we use a 9-bin color space based on HSV color representation. Three bins stand for different luminance values (black, white and gray), while other bins count the occurrences of basic color tones (red, yellow, green, cyan, magenta and pink). Input colors are transformed into HSV color space, followed by gray tone classification. This is done by analyzing the S and V color components. Pixel's color is considered to be gray if $V < 0.1 \ \vee \ S < 0.1 + \frac{0.01}{V^2}$. We define three levels of monotone illumination: black ($V \leq 0.23$), gray ($0.23 < V < 0.85$) and white ($V \geq 0.85$). After that, color classification is performed on pixels that at previous step were not classified as grayscale. The color tone is determined by splitting the H color component into 6 equally spaced regions with centers at $\{0.083, 0.25, 0.417, 0.583, 0.75, 0.917\}$ and mapping pixel's color to the closest color region. The use of a limited color description of this type accounts for the fact that slight variations in half tones are hardly detectable by the HVS in the absence of a reference image, and this information is useful only when visually very close images (near duplicates) are compared. On the other hand, the absence or presence of some basic color tones has a

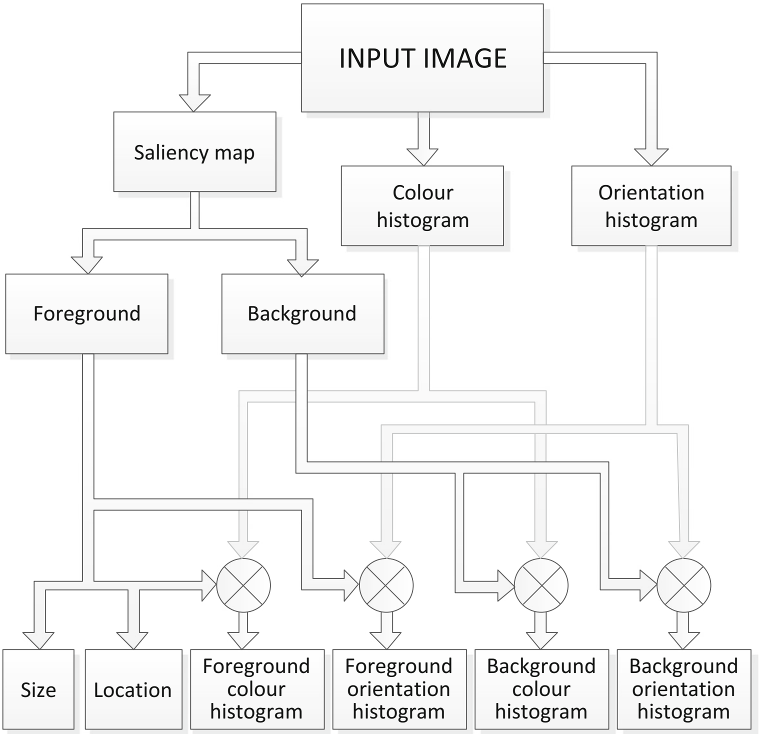

INPUT IMAGE
Saliency map
Colour histogram
Orientation histogram
Foreground
Background
Size
Location
Foreground colour histogram
Foreground orientation histogram
Background colour histogram
Background orientation histogram

Fig. 3 Feature extraction scheme: several visual features are computed for foreground and background parts of images (i.e., salient and not salient areas detected by the saliency detection method described in Section 3.1). Visual features evaluation corresponds to the second step of the proposed method and is required for the re-ranking of search results

great impact on perception. In addition, in histograms with large number of bins even tiny variation in color spectrum results in large feature distance. As reported in Fig. 3 colour histograms are computed for both foreground (salient part) and background (not salient part of the image).

Orientation histograms are also employed in our method due to several reasons. First, they allow a simple yet effective analysis of texture contents. Second, they allow estimating the observation viewpoint, in particular for objects that have a dominant orientation of straight edges on their bodies. This is the typical case for man-made objects like cars, building, etc. Orientations are detected by applying directional filters at different scales. In this work, we employ Leung-Malik (LM) filters [14], applied to first and second derivatives (of Gaussians) at 6 orientations {0°, 30°, 60°, 90°, 120°, 150°} and 3 scales. Responses at different scales and derivatives are summed up per each orientation. Also orientation histograms are computed for both foreground (salient part) and background (not salient part of the image) as shown in Fig. 3.

Finally, the saliency map allows extracting object-specific features, in particular size and location. The size is computed by normalizing the area of the foreground by the image size. The location is defined as the centroid of the foreground region, normalized over image dimensions. All the extracted features are exploited in the following step to re-rank images for search results diversification.

3.3 Search results diversification

Ranking is the key component of the system. Given a query, in order to find relevant and diversified results, it is necessary to find a suitable trade-off between similarity and diversity of images, which are controversial constrains. Since pure content-based search is still a tough problem [19], and the set of features used is insufficient to always achieve a satisfactory accuracy of visual search results, we assume to have in input a set of images returned by text-based search, providing a set of retrieved images with acceptable precision (tipically higher than 0.5). Thus, we limit our task to re-ranking of results in order to achieve a higher diversity on top N results. In principle, our system acts as a post-retrieval filter that sorts the results to increase the diversity.

As previously pointed out, the major contribution of our work is in the use of saliency to perform this task in a more effective way. This goal can be achieved in different ways: for instance, one can force foreground similarity while differentiating the background, thus resulting in the same object appearing in different contexts. On the contrary, one may differentiate the foreground independently of the background, thus achieving a larger variety of subjects. This way of proceeding however would neglect the strong correlation between foreground and background, which appears evident when analyzing the data (see Fig. 4). Another problem is that frequently occurring images should be promoted to the top places, as very rare images are likely to be less relevant. According to the above considerations, we propose a weighting method that jointly considers background and foreground diversity, while at the same time putting frequently occurring images at the top places.

Given the feature vectors associated to a pair of images im_1 and im_2 (see Section 3.2), we compute their dissimilarity according to the following equation:

$$D(im_1, im_2) = \sum_{i=1}^{n} w_i \cdot Dist\left(f_1^i, f_2^i\right) \tag{6}$$

where $Dist\left(f_1^i, f_2^i\right)$ represents the distance between image im_1 and im_2 with respect to feature f^i, w_i is the corresponding weight, and n is the number of employed features (in this work $n = 6$). Dissimilarity of histogram features is computed using cosine distance. The initial order is not used by our method. Candidate images are progressively selected by maximizing the ranking score term RS. For the sake of simplicity we used a linear ranking method:

$$RS(im) = w_{res} \cdot D_{res}(im) - w_{nran} \cdot D_{nran}(im), \tag{7}$$

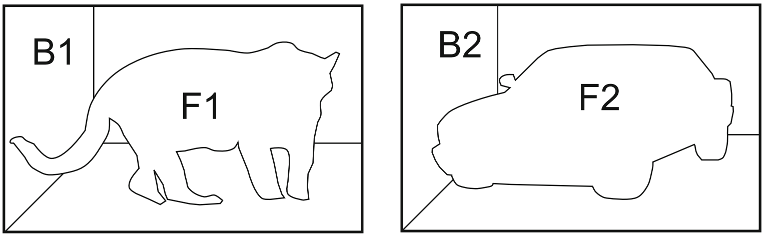


Fig. 4 Foreground (F) and background (B) correlation example. Consider the left image to be a picture of a mammal in its natural environment while the right one is a picture of a man-made object in an industrial environment. Then $P(F = F1|B = B1) \gg P(F = F1|B = B2)$ ($P(\cdot)$ denotes the probability function), likewise $P(F = F2|B = B2) \gg P(F = F2|B = B1)$

where $D_{nran}(im)$ is the overall normalized distance of image im to images in unranked results list, while $D_{res}(im)$ is the overall normalized distance of image im to images in results list. Relevant weights are w_{nran} and w_{res}.

As a result, the optimization is done by maximizing the dissimilarity with previously selected images and minimizing the dissimilarity with unranked images (see also Algorithm 1). Thereby diversity is achieved through promotion of representative images from the unranked list and penalty of similar images in the results list. This can be achieved more effectively since we treat differently information about the foreground (salient part) and background (background).

Algorithm 1 Re-ranking algorithm

Input:

- Top N results im_1, im_2, .. im_N returned by text-based search.
- Relevant weights w_{res} and w_{nran} and weights w_i corresponding to features $f^i, i = 1..n$.

Output: The re-ranked list RES of images $\{im_1, im_2, .. \, im_N\}$.

Method:

```
$RES = \varnothing$
$NRAN = \{im_1, im_2, .., im_N\}$
while $||NRAN|| > 0$ do
    for all $im_c \in NRAN$ do
        $D_{nran}(im_c) = \sum_{im_t \in NRAN \setminus \{im_c\}} \sum_{i=1}^{n} w_i \cdot Dist(f_c^i, f_t^i)$
        $D_{res}(im_c) = \sum_{im_t \in RES} \sum_{i=1}^{n} w_i \cdot Dist(f_c^i, f_t^i)$
        $RS(im_c) = w_{res} \cdot D_{res}(im_c) + w_{nran} \cdot D_{nran}(im_c)$
    end for
    $im_{max} = im_c \rightarrow max(RS(im_c))$
    $RES = RES \cup \{im_{max}\}$
    $NRAN = NRAN \setminus \{im_{max}\}$
end while
```

4 Validation

Assessing the diversity of image retrieval results is an unsolved problem so far. Therefore, in Section 4.1 we will briefly discuss possible metrics and introduce the proposed coverage measure, which will be used in Section 4.2 to evaluate the experimental results. The validation was performed on different publicly available datasets. Additionally, we will present a further test performed on the dataset collected within the MediaEval 2014 task on Retrieving Diverse Social Images. For this last test we evaluated the results on the basis of the ground-truth and metrics used in the benchmark.

4.1 Diversity evaluation

Evaluation is one of the toughest parts of the work since there is no commonly accepted metric to measure diversity in retrieval. For example, in ImageCLEF contest diversity was measured as the number of clusters in top 20 results. Such clusters were hand-designed by a group of experts. MediaEval task on image diversity used a similar approach. Although this metric seems to be reasonable, it has some disadvantages. When input text query refers to several possible semantic concepts (e.g., jaguar) clusters are naturally representing different concepts. However, when possible retrieval results are less ambiguous it is not clear how cluster centers are selected. It may occur that some clusters have different distance from each other, thus inclusion of new clusters into the ranking has different impact on percaptual diversity. Another possible approach is the use of taxonomy tree, whose leaves represent possible values of different properties. In this case, the diversity of a set of images may be computed as the number of branches covered by a set. However, this approach also has some disadvantages. First, the tree would contain redundant information with some properties repeating at different levels of the tree. In addition, the ordering of properties would affect the number of leaves in the tree, producing different diversity scores. Moreover, although recent works proposed several plausible diversity measures for tagged images, they are not fully applicable in our case. For example, for the case of the commonly used approach of data clustering, the number of clusters for a category consisting of 100 images can be as high as 70 clusters. If we take into account first 20 images retrieved, comparison of different rankings is meaningless as often these 20 images will correspond to 20 clusters. Other measures that require semantic understanding are possible but require a natural language processing framework.

In this work we propose a novel metric to evaluate diversity, trying to keep into account both semantic and visual diversity. Indeed, since saliency discriminates between foreground and background parts, it allows to achieve object representation diversification, which can be combined with concept diversity. The measure we propose is based on text-based representation of visual content by annotations. Such an annotation consists of a list of properties, encoding both visual and semantic variations of the main object within a given set. Each property consists of a list of tags that define its possible values. To each tag we assign its weight, computed as follows:

$$w_t = \frac{t_i}{i \cdot p}, \tag{8}$$

where t_i is the number of images this tag was assigned to, p is the number of properties for a set of images, and i is the total number of images in the initial image set. Diversity is measured as the coverage over tags. More details about data annotation are given in Section 4.2.

Coverage of a set is defined as the sum of weights of unique tags assigned to images in this set. Thereby only weights of newly introduced tags are counted, and the maximum possible coverage value is 1. In Fig. 5 we give an example of how coverage is computed. Depending on the number of analyzed properties, the measure can i) increase proportionally with the number of diverse properties introduced, or ii) remain steady if all properties are already present. The proposed measure allows capturing both diversity and relevance. This is done by giving higher weight to tags that are assigned to more images. Then, the overall score increases more rapidly when an image that represents a larger cluster is selected.

4.2 Experiments

For the evaluation of the approach a dataset was created based on three image datasets: Caltech 256 [7], a subset of Corel database [31], and Pascal VOC 2008 [5]. Caltech 256 and Corel provide category-based image sets. There is not any grouping in Pascal VOC 2008 dataset, but all images are provided with text description containing information about the type, bounding box and pose of objects of interest present on images. Image sets were created by applying a query-by-type on several categories. However, for some images the annotated objects were occupying a very small area, making them perceptually irrelevant.

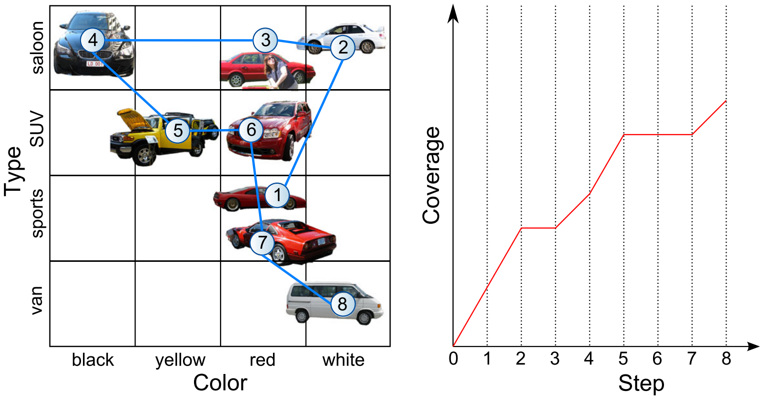
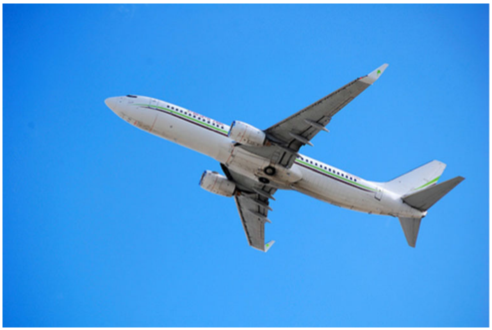


Fig. 5 Example of coverage measure computation for a category with two properties. It can be noticed that coverage measure increases only when a new value for a property is included to the ranking. For example, at step 3 the coverage does not increase since both properties are already covered (type is the same as step 2 and color is the same as step 1), while at step 4 an image with a new value (*black*) for color property is added resulting in coverage increase. When several new values appear in a single step, the increase is higher (e.g., steps 1-2 and 4-5 compared to 3-4 and 7-8)

Thus, images where the target object occupies less than 10 percent of the total area were excluded. We considered 100 images per category.

Since we created separate sets of annotations for each category, our dataset captures variations of the main visual concept and its background. After analysis of the dataset we came up with an annotation guide that encodes the following properties: colors of the main object, quantity of objects belonging to the main object class, location and size of the object, subtype of the object, viewing angle, distance, etc. These annotations provide a compact description of the visual properties of each image. Although there are plenty of other possible properties one can add, there is always a problem of subjectivity. Then, we avoided considering properties that may create ambiguity due to subjective interpretation. In addition, applicable properties are very dependent on the content, thereby only relevant properties were included into annotations for each category. The ground-truth was created by three experts. Notice that we could have come up with a fixed set of properties but it would have resulted in few properties that hardly grasp semantic and visual content of categories. Moreover, omitting irrelevant properties is equal to use a large fixed set of properties for all categories, tagging as ”null” properties that are irrelevant for a given class. An example of annotation and properties for category aeroplane (Pascal VOC 2008) are reported in Fig. 6 and Table 2, respectively.

As a first experiment, we performed the evaluation of ground-truth data. Although there are no ground-truth saliency data in all three datasets, Pascal VOC 2008 includes object segmentation ground-truth data that are very close to output maps generated by our object-wise saliency detector. Therefore, here we assume that labeled data correspond to the main object of an image. For comparison we decided to consider first 20 re-ranked images, since this is the usual set of images shown per page and it is also suitable to illustrate differences in diversity. Results are reported in Table 3, where we compared performances of the re-ranking method using the entire image area (without saliency), the ground-truth labeled data (labeled data), and the automatically extracted saliency maps (with saliency), respectively. As it can be seen, inclusion of saliency data, no matter if it is extracted automatically



Background	**Color**	**Type**	**Field**	**View**
sky	bright	jet	far field	bottom

Pose	**Number**	**Light**	**Trail**
flying	one	day	no trail

Fig. 6 Example of annotation of an image of the category aeroplane (Pascal VOC 2008)

Table 2 Example of properties for the category aeroplane (Pascal VOC 2008)

Background	airport, hangar, sky, water, land
Color	bright, dark, colourful
Type	jet, tourism, fighter, acrobatic, vintage, seaplane, toy
Field	far field, near field
View	wing, frontal, side, bottom, top, back, interior, window, cockpit
Pose	flying, park/taxi, takeoff, landing, crashed
Number	one, many
Light	day, evening, sunset, interior
Trail	white, colored, no trail

Table 3 Experiment 1: Coverage measure for first 20 images, comparing the method not exploiting saliency (without saliency), the method using ground-truth data (labeled data) and the proposed method exploiting saliency (with saliency)

Category	Without saliency	Labeled data	With saliency
Bird	0.972	0.970	0.980
Car	0.939	0.944	0.959
Cow	0.970	0.985	0.983
Boat	0.846	0.893	0.893
Sheep	0.937	0.950	0.939

or provided as a ground-truth data, improves diversity. In addition, close performance of automatically extracted maps with ground-truth data shows that the proposed detection tool gives reasonable maps, whose accuracy is sufficient for the objective. Slight difference in performance can be explained by the fact that not always labeling belongs to a visually salient object. Notice that the proposed approach depends on 8 weights ($w_i i = 1, ...6$ in (6) and w_{res} and w_{nran} in (7) while if we exclude the use of saliency for comparison we just have to set 4 weights (2 in (6) and 2 in (7)). The values used in our experiments were: $\{7.6, 1.8, 2.0, 0.6, 4.0, 0.7\}\{1, 1\}$ and $\{3.7, 5.7\}\{1, 0.7\}$, respectively. These values have been obtained based on an optimization study over the entire dataset performed by using genetic algorithms.

While the first test proves that the concept of using salicency to improve diversity is viable, the small size of the dataset and the limited availablility of ground-truth labelled data make it insufficient for a through validation. Therefore, we performed another test that covers more categories and gives more information about generalization of our approach. Results of this comparison are reported in Table 4. The evaluation was performed on 22

Table 4 Experiment 2: Coverage measure, for 20 images, for the proposed method (with saliency), the method not exploiting saliency information (without saliency) and the method exploiting a different saliency detector [17]

Category	With saliency	Without saliency	Using [17]
Airplane (Caltech 256)	0.895	0.846	0.880
Bear (Caltech 256)	0.707	0.663	0.636
Blimp (Caltech 256)	0.863	0.729	0.781
Bridge (Caltech 256)	0.712	0.521	0.564
Butterfly (Caltech 256)	0.713	0.685	0.763
Gas-pump (Caltech 256)	0.477	0.454	0.474
Pyramid (Caltech 256)	0.668	0.579	0.645
Teapot (Caltech 256)	0.615	0.586	0.485
Sea animal (Corel)	0.943	0.869	0.931
Fox (Corel)	0.915	0.832	0.924
Military vehicle (Corel)	0.835	0.668	0.838
Train (Corel)	0.995	0.942	0.995
Wolf (Corel)	0.924	0.798	0.924
Aeroplane (Pascal VOC 2008)	0.289	0.370	0.310
Bicycle (Pascal VOC 2008)	0.861	0.842	0.880
Bird (Pascal VOC 2008)	0.912	0.823	0.866
Boat (Pascal VOC 2008)	0.683	0.635	0.649
Car (Pascal VOC 2008)	0.814	0.739	0.821
Cow (Pascal VOC 2008)	0.916	0.801	0.630
Motorbike (Pascal VOC 2008)	0.389	0.386	0.395
Sheep (Pascal VOC 2008)	0.817	0.775	0.820
Table (Pascal VOC 2008)	0.840	0.753	0.703
Average	0.763	0.695	0.723

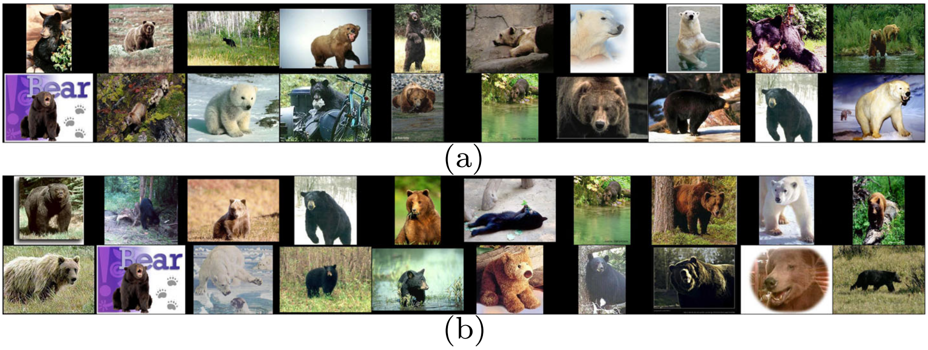


Fig. 7 Example of re-ranking for category bear using the method with saliency information (**a**) compared to the method without saliency information (**b**). Ranking (**a**) provides no near-duplicates, more instances of white bears, age variation is higher (notice images of offspring), more locations and attitudes are captured and there are pictures of a group of bears

categories taken from the three mentioned datasets. As it can be seen in Table 4 the use of saliency information results in a diversity increase of 11 %. Average coverage measure is 0.763 by using the proposed approach (with saliency), while average result without saliency is 0.695. The use of saliency allows improving performance in 21 out of 22 cases. In addition, we introduced a comparison with same method fed by a different saliency detector [17]. Category coverage using extraction method described in Section 3 and in [17] are both improving performances of the method with no use of saliency, although the method in this work outperforms the other (0.763 versus 0.723 on average). Figures 7 and 8 show visual comparison of re-rankings for category bear and pyramid respectfully. In both case it is possible to notice how the proposed re-ranking with the use of saliency information (panels

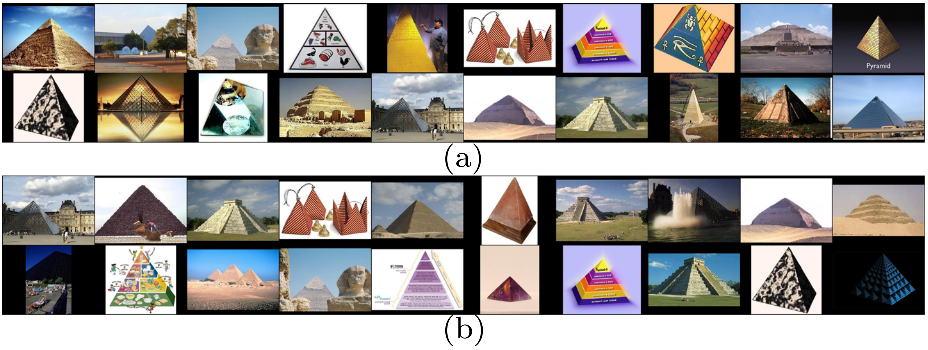


Fig. 8 Example of re-ranking for category pyramid using the method with saliency information (**a**) compared to the method without saliency information (**b**). Ranking (**a**) provides no near-duplicates, more shapes are captured and there is also an indoor picture

Table 5 Experiment 3: Results on MediaEval 2014 - Retrieving Diverse Social Image Task

	With saliency			Without saliency		
Set	P@20	CR@20	F1@20	P@20	CR@20	F1@20
Development set	0.770	0.408	0.531	0.738	0.370	0.490
Test set	0.774	0.422	0.530	0.744	0.403	0.506
All	0.773	0.419	0.530	0.743	0.396	0.503

a) outperforms the method not exploiting saliency (panels b), by avoiding near-duplicates and providing more viewpoints (in terms of perspective, position, shape, background). It is to be pointed out that in this experiments the re-ranking was applied to a set of relevant images. Therefore, we do not report precision and recall values which are equal to one.

As a final experiment, we tested our approach on the recent MediaEval 2014 "Retrieving Diverse Social Images" database (45.000 images from 153 locations spread over 35 countries allover the world) analyzing the results on the basis of the official metrics used in the task (also in this case the ground-truth was produced by a set of experts). In this task, participants received a ranked list of photos for each location retrieved from Flickr using its default "relevance" algorithm and they had to refine the results by providing a ranked list of relevant and diverse representations of the query. The evaluation metrics are computed based on the precision P, the cluster recall CR, and the harmonic mean $F1 - score$ of P and CR. These values are measured at different cut-off points and the official ranking was $F1 - score$ at the cut-off point 20: $F1@20$. In Table 5, we show the results obtained

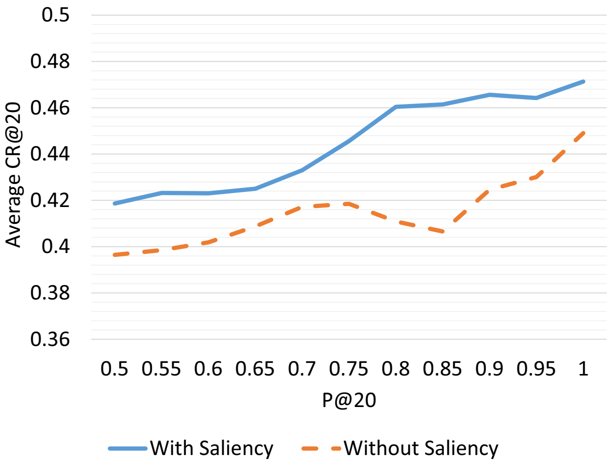


Fig. 9 Average cluster recall for increasing values of lower boundary in precision

by applying our approach on this challenging dataset, both for the Development set (30 locations) and the Test set (123 locations). We have used here the same weights as in the previous experiments, optimised for the original dataset. This demonstrates that the method is able to generalize to different datasets. Notice that, since our method was designed as a re-ranking algorithm operating on sufficiently relevant results, we have discarded locations where the precision is below 0.5. As it can be seen, the use of saliency information results in the official score increase of 5.4 % on average. In Fig. 9 we report the average cluster recall at the cut-off point 20 for increasing values of lower boundary in precision (i.e., filtering out all cases with lower precision). It can be noticed that the improvement is particularly high when the precision is above 0.70-0.75, which are typical precision values in many common application scenarios. In fact, in this situation the framework may fully express its diversification capability over a wide set of relevant results.

5 Conclusions

In this paper we described how saliency information can be applied for the task of diversification of retrieval results. The evaluation on different datasets showed that both conceptual and perceptual diversity can benefit from incorporating visual saliency information. The proposed approach can be easily extended by adding other visual features to include further dimensions of the diversity.

Further work will include the analysis of semantic features either as text-based or bag-of-visual words information and the incorporation with visual saliency information. This could lead to improved discrimination of the content to further increase the diversity.

Acknowledgments This work was supported by the Deanship of Scientific Research of the King Saud University through the International Research Group under Project IRG14-20.

References

1. Carbonell J, Goldstein J (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
2. Clarke C, Craswell N, Soboroff I, Ashkan A (2011) A comparative analysis of cascade measures for novelty and diversity. In: ACM International Conference on Web search and data mining
3. Dang-Nguyen DT, Piras L, Giacinto G, Boato G (2014) Retrieval of diverse images by pre-filtering and hierarchical clustering. In: MediaEval Benchmarking Initiative for Multimedia Evaluation
4. Deselaers T, Gass T, Dreuw P, Ney H (2009) Jointly optimising relevance and diversity in image retrieval. In: ACM International Conference on Image and Video Retrieval
5. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2008) The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results, http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html
6. Gelasca E, Tomasic D, Ebrahimi T (2005) Which colors best catch your eyes: a subjective study of color saliency. In: International Workshop on Video Processing and Quality Metrics for Consumer Electronics
7. Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology, http://authors.library.caltech.edu/7694

8. Hoiem D, Efros A, Hebert M (2005) Geometric context from a single image. IEEE International Conference of Computer Vision
9. Huang J, Yang X, Fang X, Lin W, Zhang R (2011) Integrating visual saliency and consistency for re-ranking image search results. IEEE Trans Multimed 13(4):653–661
10. Ionescu B, Popescu A, Lupu M, Ginsca AL, Mueller H (2014) Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation. In: MediaEval Benchmarking Initiative for Multimedia Evaluation
11. Kennedy LS, Naaman M (2008) Generating diverse and representative image serach results for landmarks. In: ACM International Conference on World Wide Web
12. Koehler K, Guo F, Zhang S, Eckstein MP (2014) What do saliency models predict. J Vis 14(3): 1–27
13. van Leuken RH, Garcia L, Olivares X, van Zwol R (2009) Visual diversification of image search results. In: ACM International Conference on World Wide Web
14. Leung T, Malik J (1999) Recognizing surfaces using three-dimensional textons. In: International Conference on Computer Vision, pp 1010–1017
15. Li H, Tang J, Li G, Chua TS (2008) Word2image: Towards visual interpreting of words. In: ACM International Conference on Multimedia
16. Li J, Ji Z, Zhang J, Su YT (2012) Generating diverse and relevant image searching results with divrank. In: International Conference on Machine Learning and Cybernetics
17. Liu T, Yuan Z, Sun J, Wang J, Zheng N, Tang X, Shum HY (2011) Learning to detect a salient object. IEEE Trans Pattern Anal Mach Intell 33(2):353–367
18. Muratov O, Zontone P, Boato G, De Natale FGB (2011) A segment-based image saliency detection. In: IEEE International Conference on Acoustics, Speech and Signal Processing
19. Paramita ML, Sanderson M, Clough P (2009) Diversity in photo retrieval: Overview of the imageclefphoto task 2009. In: International Conference on Cross-language evaluation forum: multimedia experiments
20. Ramanathan S, Katti H, Sebe N, Kankanhalli M, Chua TS (2010) An eye fixation database for saliency detection in images. In: European Conference on Computer Vision
21. Ramanathan S, Yanulevskaya V, Sebe N (2011) Can computers learn from humans to see better?: Inferring scene semantics from viewers' eye movements. In: ACM International Conference on Multimedia, pp 33–42
22. Riche N, Duvinage M, Mancas M, Gosselin B, Dutoit T (2013) Saliency and human fixations: State-of-the-art and study of comparison metrics. In: IEEE International Conference on Computer Vision
23. Rudinac S, Hanjalic A, Larson M (2013) Generating visual summaries of geographic areas using community-contributed images. IEEE Trans Multimed 15(4):921–932
24. Sanderson M, Tang J, Arni T, Clough P (2009) What else is there? search diversity examined. In: European Conference on Information Retrieval
25. Shi R, Liu Z, Du H, Zhang X, Shen L (2012) Region diversity maximization for salient object detection. IEEE Signal Proc Lett 19(4):215–218
26. Song K, Tian Y, Gao W, Huang T (2006) Diversifying the image retrieval results. In: ACM International Conference on Multimedia
27. Tollari S, Mulhem P, Ferecatu M, Glotin H, Detyniecki M, Gallinari P, Sahbi H, Zhao ZQ (2009) A comparative study of diversity methods for hybrid text and image retrieval approaches. In: Workshop of the Cross-Language Evaluation Forum
28. Wang M, Yang K, Hua XS, Zhang HJ (2010) Towards a relevant and diverse search of social images. IEEE Trans Multimed 12(8):829–842
29. Yu Z, Lu L, Guo Y, Fan R, Liu M, Wang W (2014) Content-aware photo collage using circle packing. IEEE Trans Vis Comput Graph 20(2):182–195
30. Zontone P, Boato G, De Natale FGB, De Rosa A, Barni M, Piva A, Hare JS, Dupplaw D, Lewis PH (2009) Image diversity analysis: context, opinion and bias. In: Living Web Workshop
31. Zontone P, Boato G, Hare J, Lewis P, Siersdorfer S, Minack E (2010) Image and collateral text in support of auto-annotation and sentiment analysis. In: Workshop on Graph-based Methods for Natural Language Processing, pp 88–92
32. Zwol RV, Murdock V, Pueyo LG, Ramirez G (2008) G.: Diversifying image search with user generated content. In: ACM International Conference on Multimedia Information Retrieval

Giulia Boato is Assistant Professor at the Department of Information Engineering and Computer Science (DISI) of the University of Trento (Italy) and professor of the courses Digital signal processing and Data hiding within the M.Sc. Degree in Telecommunications Engineering. From 2008 to 2011 she has been coordinating the Multimedia Signal Processing and Understanding Lab. She was in the Project staff of many projects (FP7 FET-IP LIVINGKNOWLEDGE, FP7 IP GLOCAL, FP7 CA ETERNALS). She is co-chair of the International Workshop Living Web: making diversity a true asset (Washington DC, October 2009) within the International Semantic Web Conference 2009 and of the workshop on Event-based Media Integration and Processing co-located with ACM Multimedia conference 2013. She is reviewer for many international journals, e.g., IEEE Transactions on Information Forensics and Security, IEEE Transactions on Signal Processing, IEEE Transactions on Multimedia, IEEE Transactions on Circuits and Systems for Video Technology. Her research interests are focused on image and signal processing, with particular attention to multimedia data protection, data hiding and image forensics, but also intelligent multidimensional data management and analysis.

Duc-Tien Dang-Nguyen received the Ph.D. degree from the Department of Information Engineering and Computer Science (DISI) of the University of Trento (Italy). In 2014, he has been a post-doctoral research fellow at Pattern Recognition and Applications Lab (PRa Lab), Department of Electrical and Electronic Engineering (DIEE) of University of Cagliari. His main interests include multimedia forensics, multimedia retrieval, and multimedia event analysis.

Oleg Muratov did the PhD at the Department of Information Engineering and Computer Science (DISI) of the University of Trento (Italy). His thesis focused on the development of new visual saliency detectors and their application to the problem of image diversification.

Naif Alajlan is Associate Professor, Computer Engineering Dept., King Saud Univ., Riyadh. He is member of PAMI research group of the University of Waterloo since 2003. He is reviewer for many international journals, e.g., IEEE Transactions on Signal Processing, IEEE Transactions on Image Processing, Image and Vision Computing. His research interests spam from pattern recognition, to image processing and are focused in particular on shape representation and matching and semantic image retrieval.

Francesco G. B. De Natale is Full Professor of Telecommunications Engineering (from 2003). He has been the Head of the Department of Information Engineering and Computer Science (DISI) from 2006 to 2009 and currently leads the Research Lab on Multimedia Communications (mmlab.disi.unitn.it) and the MMSPI (Multidimensional Multimodal Signal Processing and Interpretation Lab) of the Italian branch of the European Institute of Technology (EIT-ICTLabs@Italy). His research interests are focused on multimedia communications, with particular attention to multidimensional signal processing, analysis, storage and retrieval. His results are witnessed by the publication record, with more than 50 works published on journal articles and more than 100 peer reviewed international conference papers. He was General Co-Chair of the Packet Video Workshop (PV-2000), Program Co-Chair of the IEEE Intl. Conf. on Image Processing (ICIP-2005), and General Chair of the ACM Intl. Conf. on Multimedia Retrieval (ICMR-2011). He has been Associate Editor of the IEEE Trans on Multimedia and of the IEEE Trans. on Circuits and Systems for Video Technologies, as well as a member of the IEEE Signal Proc. Society Technical Committee on Multimedia Signal Processing (MMSP). Currently, he is member of the Board of Directors of the Italian Consortium for Telecommunications (CNIT), and the UniTN delegate in the Assembly of the EU EIT ICTLabs.