Image and Vision Computing 28 (2010) 927–939

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

# Learning natural scene categories by selective multi-scale feature extraction

Alessandro Perina [a,*], Marco Cristani [b,1], Vittorio Murino [b,2]

[a] Dipartimento di Informatica, University of Verona, Strada Le Grazie 15, 37134 Verona, Italy
[b] IIT, Italian Institute of Technology, Via Morego 30, 16163 Genova, Italy; Dipartimento di Informatica, University of Verona, Strada Le Grazie 15, 37134 Verona, Italy

ARTICLE INFO

*Article history:*
Received 5 January 2009
Received in revised form 16 November 2009
Accepted 19 November 2009

*Keywords:*
Image representation
Image classification
Generative modeling

ABSTRACT

Natural scene categorization from images represents a very useful task for automatic image analysis systems. In the literature, several methods have been proposed facing this issue with excellent results. Typically, features of several types are clustered so as to generate a vocabulary able to describe in a multi-faceted way the considered image collection. This vocabulary is formed by a discrete set of visual codewords whose co-occurrence and/or composition allows to classify the scene category. A common drawback of these methods is that features are usually extracted from the whole image, actually disregarding whether they derive properly from the natural scene to be classified or from foreground objects, possibly present in it, which are not peculiar for the scene. As quoted by perceptual studies, objects present in an image are not useful to natural scene categorization, indeed bringing an important source of clutter, in dependence of their size.

In this paper, a novel, multi-scale, statistical approach for image representation aimed at scene categorization is presented. The method is able to select, at different levels, sets of features that represent exclusively the scene disregarding other non-characteristic, clutter, elements. The proposed procedure, based on a generative model, is then able to produce a robust representation scheme, useful for image classification. The obtained results are very convincing and prove the goodness of the approach even by just considering simple features like local color image histograms.



## 1. Introduction

In the machine learning literature, the term "scene" is commonly defined as *a semantically coherent, nameable human-scaled view of a real world environment* [1]. The capability of analyzing and classifying accurately an imaged scene is highly useful for automatic image analysis systems in a wide variety of tasks. Other than the pure classification of the environment contained in a picture, individuating the scene category may help also in object recognition, providing a context on the possible semantic labels of the objects identities (e.g., a shark is rare to see in a mountain environment) [2]. Conversely, as reported in a recent work [3], a human being does not need to perceive the objects in a scene to identify its semantic category: behavioral and computational studies show that humans rely on global visual properties to exploit scene classification, instead of performing recognition of particular objects in a scene.

In a computational context, scene categorization methods can be partitioned into two parts: local and global methods. Such distinction follows behavioral studies on human perception, where local and global mechanisms are supposed to work during the early stages of image acquisition. Nowadays, the relative importance of each paradigm is unclear and represents an open research topic worth to be further investigated. Local methods extract from the image a set of unordered local descriptions, pooling them together and building a classifier in which the global structure of the image is usually lost. The most known and used paradigm belonging to this category is the so-called "bag of words" [4,5]. On the other side, global methods use information generated through the presence of large spatial structures, and the spatial arrangement of lighter and darker areas in an image [6]. In these methods global features are composed by local patterns in which their relative spatial layout is preserved and learnt. Therefore, such methods, first, find global spatial structures in an image, and, second, extract local descriptions that explain more in detail the spatial layout [3].

Each class of methods brings its own pros and cons. Recently, in [7], an interesting experiment was proposed, in which local and global ways to perform scene categorization were analyzed in a comparative way, measuring the classification accuracy of human subjects against automatic systems. This study showed, as expected, that human classification outperforms automatic classification. More interesting, global and local methods are discovered to be more effective in particular cases: rivers, lakes, and mountains are categorized better using global information, whereas coasts,

* Corresponding author. Tel.: +39 045 8027803; fax: +39 045 8027068.
*E-mail addresses:* alessandro.perina@univr.it (A. Perina), marco.cristani@univr.it (M. Cristani), vittorio.murino@univr.it (V. Murino).
[1] Tel.: +39 045 8027988.
[2] Tel.: +39 045 8027996.

0262-8856/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2009.11.007

forests and plains are categorized better using local information [7]. Like for human beings, it has been shown that a hybrid classifier formed by a joint global and local response outperforms a single classifier, suggesting that both global and local mechanisms cooperate to categorize an image. Anyway, both schemes do not work properly (at best of their performances) when the images analyzed present objects not being characteristic of a particular scene; for example, faces or persons in foreground are important causes of misclassification in such methods.

For convenience, we term as background (BG) the scene we want to categorize, and with foreground (FG) every object which does not belong to the scene, in the sense that it does not help to intuitively assigning a precise natural scene class label to an image. We stress this definition in order to ease the understanding of the "foreground": it does not represent something that is the nearest nameable entity with respect to the camera, but its meaning here resembles the idea of foreground in the video surveillance field in which the FG is whatever unexpected, atypical for the scene observed [8]. As a matter of fact, several methods for scene categorization work with images in which no FG objects are present like [9–11]. Therefore, the task of separating FG objects from their BG environments is necessary to perform a better scene categorization.

### 1.1. Overview of the proposed approach

In this paper, we introduce a multi-scale method for image representation that, equipped with a novel generative model, leads to an intuitive natural scene classification scheme. The method is based on a novel joint global and local paradigm for feature extraction, in the sense that a global feature extraction method is possibly applied not only to the whole images, but also to different, local, image regions. Another novelty of this approach lies at the feature extraction level, carried out selectively in the image to account for useful information functional to scene background categorization.

The sketch in Fig. 1A describes the fundamental target of our approach: the idea is to disregard in an automatic fashion atypical objects in a scene, automatically pruning them as outliers away from a consistent distribution of features that genuinely encodes a natural scene. At the end, we obtain a grid representation of the image, whose cells are labelled with different natural concepts, disregarding FG occluders. Our approach is built as a multi-scale framework: at the first, highest, level the whole image is considered extracting information from all the image pixels. In the subsequent levels, only parts of the images are considered; those defined by foreground objects. We call the process of separating the background from the foreground as *BG distillation* (Section 3). At the first level, the BG distillation individuates and assigns a label to *global aspects* or *patterns* belonging to the BG, like general natural patterns like sky, water, rock, grass, etc. In this phase, FG objects are disregarded: for instance, running people on the streets, or a small bounded presence of mountains over a lake, or ducks on a lake do not concur to form BG global patterns (see Fig. 1A).

In the subsequent phase, named *BG specialization* (Section 4) the content of the images is evaluated, calculating how well a previously built global pattern can describe the image content.

Images not well described with a single global pattern are then subdivided into a set of regular non-overlapping rectangular windows, that we call *sectors* (Fig. 1B), representing a local (finer) level of analysis. After the partitioning, two different situations may hold for a sector:

1. the *sector* is described by one of the BG patterns found at the global level, and thus it can be labelled accordingly
2. the *sector* represents a novel "sector" BG pattern that is present locally with high frequency in several images.

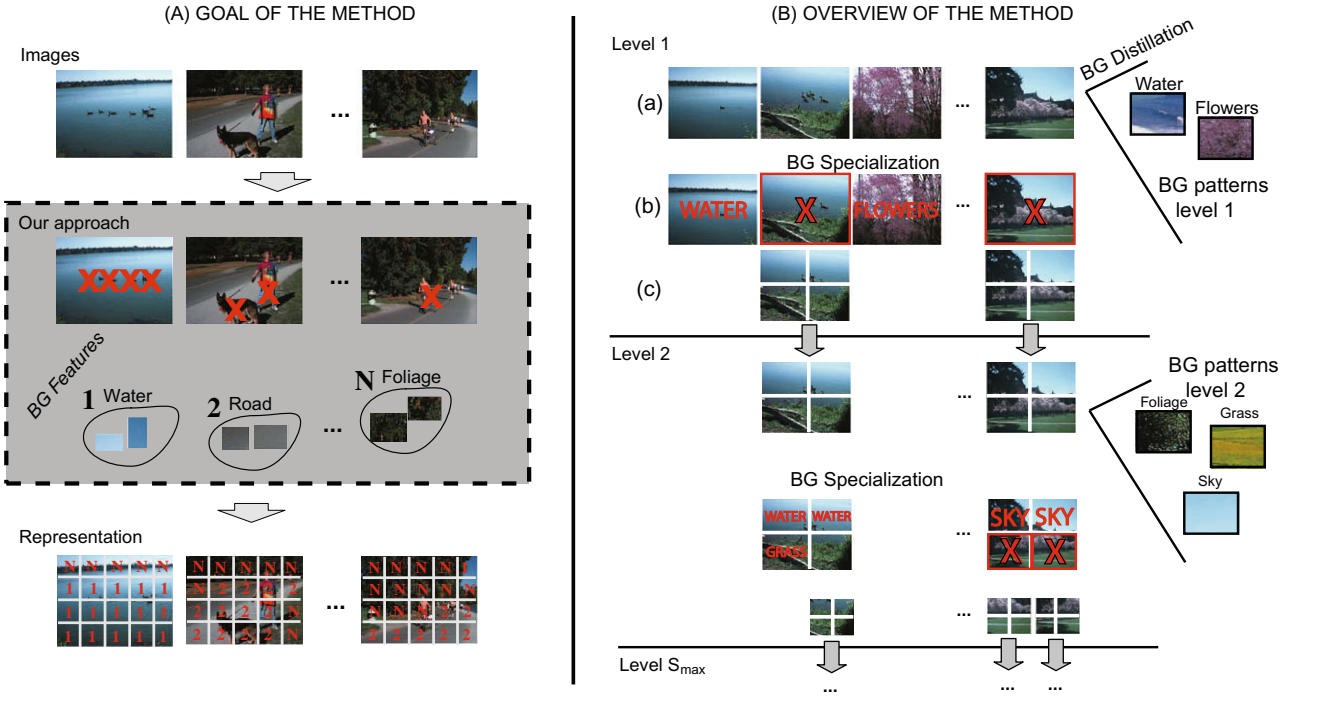


**Fig. 1.** (A) Goal of the approach: our method considers the images jointly, building thus a consistent definition of BG. This permits to obtain features that portray genuinely natural scenarios, disregarding FG objects (red crosses on the images). Such features will serve to annotate locally images with BG labels, in an automatic fashion. (B) Overview of the proposed method: during the first distillation step (a), images in their entirety are "distilled" to find typical constitutive elements (BG patterns) disregarding the FG elements. In the successive specialization step (b), the degree of fitting of each image to any BG pattern is evaluated. Images not well represented by a BG pattern are divided in four sectors (c) and the BG distillation/specialization process is repeated at a finer level. This process continues iteratively until the lowest level is reached, paying attention that, at each specialization step, the images are compared with all the previously found patterns.

In the latter situation, the sector BG patterns are extracted in the same fashion as the global BG patterns, hence, even in this case, the BG distillation permits to obtain sector BG patterns deprived of FG artifacts (Fig. 1B). Our two-step process (BG distillation and specialization) continues iteratively until a smallest patch size is considered.

At the end of the process, a grid representation of the image is produced (see Fig. 1A, bottom) by considering the smallest sectors obtained, possibly dividing larger sectors and propagating accordingly the corresponding labels.

It is worth noting that the process of BG distillation is cast in a generative framework, permitting to manage in a formal way the uncertainty derived from the BG pattern estimation (Section 3).

The rest of the paper is organized as follows. Section 2 presents a description of the state of the art. In Sections 3 and 4, the proposed framework is described. Section 5 details experimental comparisons, and, finally, the contributions of the work are summarized in Section 6.

## 2. State of the art

A widely-used taxonomy considers the methods for scene categorization as separated in *local* and *global* approaches.

### 2.1. Local methods

The main hypothesis underlying the local approaches is that a landscape depicted at different view-angles and lighting conditions produces images which are globally very different, but locally similar. This is because features which characterize natural images are very redundant, co-occurrent and therefore robust to clutter, spatial displacements, and occlusions.

Local methods have become more important in recent years, due to a successful translation of the "bag of words" paradigm [12] into the image domain [13,14,5,15,16]. Bag of words is a representation model applied originally to the text classification domain, relying on the high discriminative power of some words and on the redundancy of the language in general. The idea is to use as text descriptor the histogram of words that appeared more frequently [17]. Transferred to the image domain, a bag of words becomes a bag of "visterms" (BOV), which are local visual features co-occurring in the image. The drawback of these methods is that such representation contains no information about the spatial layout of the visterms, at the same fashion the bag of words text representation removes the words ordering information. This ambiguity generates undesirable effects of *polysemy* (one description for several images) and *synonymy* (several descriptions for one image). This issue has been faced effectively using probabilistic latent semantic analysis (PLSA) [18]. Basically, PLSA introduces an intermediate representation called *theme*, *topic*, or *concept*, which is a robust representative of several co-occurring visterms, solving also the sparseness problem of the bag-of-words paradigm. Therefore, an image can be thought of as a weighted mixture of themes [5,15,16]. In this way, bags of visterms describing a unique image become now associated with high probability to very few concepts (synonymy is minimized). In the same way, polysemy of a loosely descriptive visterm now is overcome by the concept, which is more expressive since it is conditionally linked to several visterms.

Bag of visterms-related methods in computer vision are originally utilized for object recognition, due to the fact that objects present less aspect variability than natural scenes [19].

This trend is changed recently. In [20], a local method based on bag of visterms and PLSA is proposed to perform scene categorization. Visterms here are cluster centroids of SIFT features [21], found by *K*-means [22]. Experiments are divided in three separate problems, aimed at distinguishing indoor/outdoor scenes, which are the super-ordinate-level categories defined in [23], city/landscape and indoor/city/landscape scenarios [24].

A similar approach has recently been proposed by [5]: here, a set of features extracted with four different policies (on evenly sampled grid, by random sampling the locations, using Kadir & Brady saliency detector and SIFT descriptors) was clustered by *K*-means, resulting in a set of quantized visterm. Unlike the previous approach, the image categories, other than the themes, are modelled as random variables. In this fashion, the extraction of themes is conditioned on the label of the category chosen.

In [25], multi-class SVMs were trained on the BOV representation of the member images of each scene category, where the visterms are salient points detected by difference of Gaussian operators [21]. Here, a deep analysis has been carried out on how the change of the number of different visterms affects scene classification performances.

In all the local approaches, a big effort is spent in choosing a good set of low-level features: actually, a possible inclusion of loosely representative information represents a serious drawback for all the subsequent analysis. Scene categorization is particularly sensitive to this issue: the risk is to focus only on objects accidentally appearing in a scene, disregarding the scene itself. The problem of choosing a good visterms codebook for scene categorization is, among others, addressed in [26,27,25]. In [26], the solution was to consider features well-suited for capturing natural image statistics at local level, i.e., the Weibull-based features [28].

Another contribution towards robustness of features is given in [29]. In this work, invariance with respect to affine transformations is achieved by treating the entire image dataset with an affine invariant preprocessing procedure.

The bags of visterms has been augmented with local spatial modelling in [14], including the "doublets", i.e., features formed by pair of spatially local co-occurring visterms. Spatial layout analysis techniques [30–33] are local approaches that learn: (1) the locations associated to *topics* in the images [31] and (2) the locations of the *visterms of a single topic* grouping them in a single cluster [30,32], also introducing robust management of the clutter [33].

Another approach that organizes local features in a more general spatial structure is the one proposed in [34]; the idea is to repeatedly subdividing the image and computing histogram of local features at increasingly finer resolution. This spatial pyramidal structure permits to flow down from a global point of view to a local analysis and is found to be perceptually effective. This approach has to be differentiated from the one presented in [35], where histograms are iteratively calculated on the image at different resolutions, but with a fixed number of bins.

### 2.2. Global methods

We define the class of global approaches as the one formed by methods that explicitly use information of all the pixels in the image, without eliminating or highlighting some local parts.

The use of global analysis to perform scene categorization represents the first strategy adopted in the machine learning literature [36–38].

However in the last 5 years, local methods appeared to be more effective in image analysis than global paradigms, due to the bag of visterms approaches. Anyway, the absence of an explicit and structured spatial layout description discourages a purely local approach for scene categorization applications.

An example of a global method is the approach proposed in [39]. It first performs multiple partitions of the considered images by segmenting them using Normalized Cut [40], using different parameterizations. Then, it extracts relevant invariant features from each image and visual codewords are thus obtained by

clustering the features using *K*-means. Finally, it discovers robust descriptions of segments by Latent Dirichlet Allocation [4], which is an extension of PLSA.

Oliva and Torralba propose a more general global algorithm for describing images [41], which has been refined recently in [3]. The key concept is the *spatial envelope*, which encodes five global properties of the scenes (naturalness, openness, roughness, expansion, ruggedness). The approach extracts features from the power spectrum of the images by convolving them with Gabor-like filters at 12 orientations and 5 scales. Because many filters are needed to cover the spectrum, they extract only the first 16 principal components of the images as determined by Principal Component Analysis on the training set. In the same paper, results of previous perceptual experiments are reported, which encourage the use of global methods in the scene categorization. In particular, it is discussed that human beings avoid to perform object recognition when the goal of the categorization regards the correct labelling of real environments.

In relation to the local-global taxonomy, our approach presents global and local elements. Actually, it parses images in a multiscale fashion, and most importantly, provides a finer representation of the image only when necessary, exploiting the fact that in natural scene classification, some categories are better classified with global methods, while other ones are best modelled with local methods.

## 3. Background distillation

### 3.1. Technical preliminaries: occlusion model

The BG distillation step takes inspiration from a probabilistic generative model named here occlusion model, which was proposed in [42]. The occlusion model has been applied in a toy scenario, i.e., with images that have been generated from background visual scenes, upon which one or more foreground objects have been superimposed. BG scenes and FG objects come from a set of predetermined classes; for each image, one FG class generates one instance of a particular object (a particular face) which is placed in a given position. (Fig. 2A). The goal of this model is to estimate the appearance of all the FG objects and all the BG scenes, as a set of FG and BG classes, separating them through transparency masks. An improved version of the occlusion model was proposed in [43], where the FG objects belonging to a given class were not constrained to appear in a fixed position, permitting a finite set of translations and rotations. These two generative models are meaningful and intuitive, because they express an image partition scheme which has a clear semantic interpretation (one or more foreground objects imposed on a scene). Our goal is to employ the idea of superposition of FG entities over a BG scenario in a scene classification context. Taken in their original version, none of the two models above can be applied. First, while it is intuitive in a natural scene classification framework to have a well defined finite set of natural BG scenario classes [23], the goal of estimating all the possible classes of FG patterns cannot be fulfilled, for both intuitive and technical reasons (e.g., how many FG classes?). Another lack of the two models regards robustness: other than translation and rotations, invariance with respect to scale transformations should be faced, as well as to illumination changes.

Our solution consists in proposing a novel generative model, named *occluded background* (OB) model. Such model describes images at different levels, following a quad tree structure, in order to obtain invariance with respect to translations, rotations, occlusion and scale. As a robust global image descriptor, we choose a quantized color-histogram in HSV space. In this way, pixel intensities $\mathbf{z}^{(t)} = \{z_1^{(t)} \dots z_K^{(t)}\}$ with $K$ being the number of pixels in the $t$th image, $t = 1, \dots, T$, are replaced by the histogram's bin values $\mathbf{h}^{(t)} = \{h_1^{(t)}, \dots h_H^{(t)}\}$, where $H$ varies depending on the bin quantization. Note that, in principle, whatever feature that can be described with histograms can be employed. Developments of our framework using different kind of features is subject of future work.

### 3.2. Occluded background model

The occluded background model (Fig. 2B) is essentially a mixture model (with $b$ as mixture variable) with a binary feature selection variable $m = \{m_1, \dots m_H\}$ which aims at individuating the salient observed features of each image that characterize a particular mixture component. Differently from [42], the observations are the bin values of the histogram $h = \{h_1, \dots h_H\}$ and the key structural element in our model consists in the conditional link from the mask random variable $m$ to the BG class $b$ and the presence of a dummy class which models all the FG objects, so that an histogram bin $h_i$ is generated starting from the $b$th prototype if that bin is salient, otherwise from the dummy class.

The mixture variable $b$ clusters the data in $B$ components or classes, whose centroids are called here *prototypes* and are represented by HSV histograms. The prototype represents the formal translation of the BG pattern expressed in Section 1.1. Note that $h, b, m$ are replicated for $T$ times, as expressed by the plate notation in Fig. 2B, giving rise to $\{h^{(t)}, b^{(t)}, m^{(t)}\}_{t=1,\dots,T}$.

In this way, assuming only a unique, global level of analysis (i.e., each image is an atomic entity), the generative process that forms an observed image histogram is shown in Fig. 2C and described below:

1. Choose a prototype class $b^{(t)}$ for the $t$th image, where $b^{(t)} \in \{1, \dots, B\}$. $B$ is the total number of background prototypes.
2. For a particular image histogram generated by the $b^{(t)}$th prototype, determine which bins are significant given that prototype. This is done by choosing the image binary mask $m_i^{(t)}$, with $m_i^{(t)} \in \{0, 1\}$, where $m_i^{(t)} = 1$ indicates that the $i$th bin is a salient background bin, so contributing to form the $b^{(t)}$th background class prototype.
3. Finally, the values of the histogram bin are chosen independently, given the mask, the prototype class and the noise class.

This generative process leads to the joint distribution[3]

$$P(h, m, b) = P(b) \cdot \left( \prod_{i=1}^{H} P(h_i | m_i, b) \right) \cdot \left( \prod_{i=1}^{H} P(m_i | b) \right) \quad (1)$$

In this equation, $P(h_i|m_i, b)$ can be further factorized by noticing that, if $m_i = 0$, the pixel values are generated by the foreground dummy class, whereas if $m_i = 1$ the pixel values are generated by the $b^{(t)}$th BG class so

$$P(h_i|m_i, b) = P(h_i|b)^{m_i} P(h_i|n)^{1-m_i} \quad (2)$$

We parameterize the probability $P(b)$ of a background class $b$ by $\pi_b$, which is a $B$-dimensional array representing a multinomial prior probability.

The probability that $m_i = 1$ given that the background class is $b$ (i.e., $P(m = 1|b)$), follows a binomial distribution parameterized by $\alpha_{bi}$. We call $\alpha_b = \{\alpha_{b1}, \dots \alpha_{bH}\}$ the peculiarity map, since it highlights the distinguishing bins of a particular background class providing a prior on $m_i$. In practice $\alpha_b$ could be considered a weight that codifies our certainty about how much a bin is salient for a particular class. Since the prior probability that $m_i = 0$ is $1 - \alpha_{bi}$, we can write

[3] For the sake of clarity, in the following we omit the index $t$, when not strictly necessary.

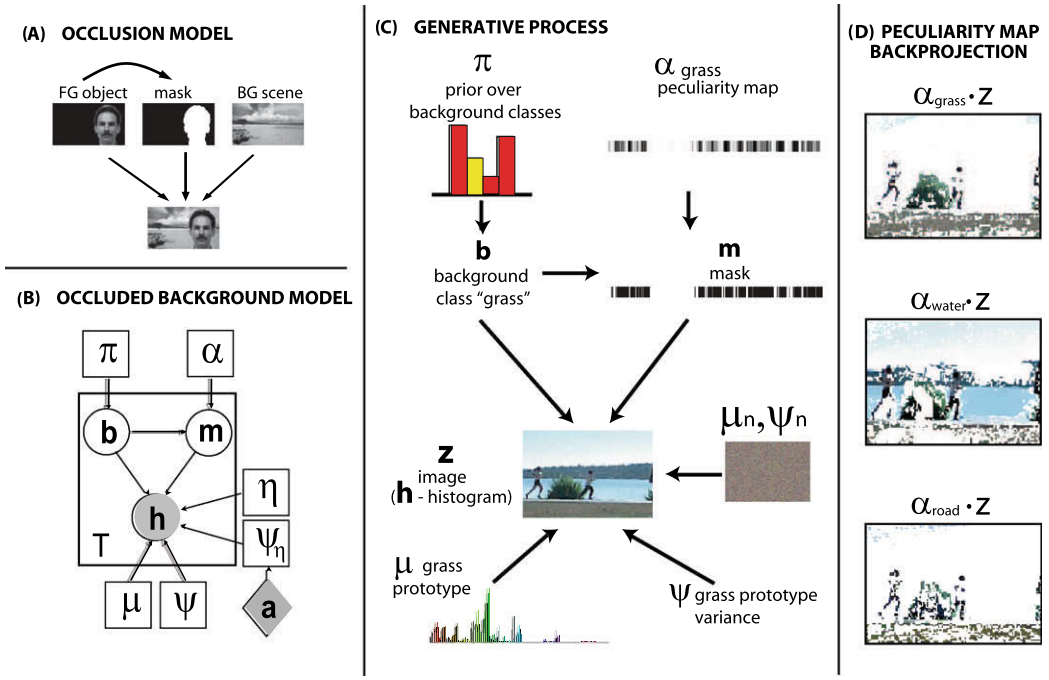


**Fig. 2.** (A) The generative process underlying the *occlusion model* [42]. (B) Graphical representation of the *occluded background model*. Circles represent the hidden variables, shaded circles are visible variables, squares are the parameters and diamonds represent constants. All the variables in the central plate are replicated $T$ times, one for each image. (C) The sketch of the generative process of the occluded background model. The example reported here uses information collected during the experimental phase. We are considering the "grass" prototype: please note that its peculiarity map has white values (=high saliency) in correspondence with the bins which model the green color (the green bins). (D) Sample images under different saliency masks, visualized via backprojection (see Section 3.3).

$$P(m_i|b) = \alpha_{bi}^{m_i}(1-\alpha_{bi})^{1-m_i} \tag{3}$$

Each bin value is modelled by a Gaussian function with parameters $\mu$ and $\psi$, so $h_i \sim \mathcal{N}(\mu, \psi^2 \cdot I)$, $I$ being the $H \times H$ identity matrix.

Each cluster centroid is hence modelled by a Gaussian of parameters $\mu_b, \psi_b$ for the background classes; $\mu_n, \psi_n$ are the respective parameters for the dummy (FG) class.

Since the dummy class $n$ models all the possible FG objects, we have to discourage low values of $\psi_n$. In other words, we have to ensure a low specificity, since FG objects are highly variable entities.

To this end, we place an inverse-gamma prior Inv-$\Gamma(\psi_n; a_1, a_2)$ on $\psi_n$ with hyper parameters $A = a_1, a_2$. This prior is used to keep $\psi_n$ within 60–100% of the dynamic range of the training data.

The parameterized version of the joint probability of the OB model is thus

$$P(h,m,b) = \pi_b \cdot \text{Inv-}\Gamma(\psi_n; a, b) \prod_{i=1}^{H} \alpha_{bi}^{m_i}(1-\alpha_{bi})^{1-m_i} \cdot \mathcal{N}(h_i; \mu_{bi}, \psi_{bi})^{m_i} \cdot \mathcal{N}(h_i; \mu_n, \psi_n)^{1-m_i} \tag{4}$$

were $\mathcal{N}(h_i; \mu_i, \psi_i)$ is the Gaussian density function calculated in $h_i$ with mean $\mu_i$ and variance $\psi_i^2$.

### 3.2.1. Free energy of the occluded background model

The model is learnt using a variational version of the Expectation–Maximization algorithm (EM) [44,45]. Here, we derive the inference ($E$ step) and the parameters update rule ($M$ step).

In this case, the only visible variable is $V^{(t)} = \{ h^{(t)} \}$; $t = 1 \dots T$, being $T$ the number of images under analysis; the hidden variables are $H^{(t)} = \{ b^{(t)}, m^{(t)} \}$. Each observation $h^{(t)}$ has a separate component variable $b^{(t)}$ and an array of $H$ masks variables $m^{(t)}$, the parameters $\theta = \{\alpha, \mu, \psi\}$ are shared among the observations, the inverse gamma prior is fixed and not learned from the data.

A standard criterion to optimize when fitting such (graphical) models is the likelihood or the log likelihood of the observed data $\log V$, obtained by summing or integrating over the hidden variables $H^{(t)}$ for a given set of parameters $\theta$. For occluded background model we have an exponential number of configurations ($2^{H \cdot T}$) so approximate methods such as variational approximations must be used. Variational approximations are based on an alternative cost, named free energy for its similarity with the quantity used in statistical physics. The free energy is defined as (re-including here the index $t$ for clarity)

$$F = -\ln P(\theta) + \sum_t \left( \sum_{b^{(t)}} \sum_{m^{(t)}} q(b^{(t)}, m^{(t)}|\theta) \log(q(b^{(t)}, m^{(t)})|\theta) - \sum_{b^{(t)}} \sum_{m^{(t)}} q(b^{(t)}, m^{(t)}|\theta) \log P(h^{(t)}, b^{(t)}, m^{(t)}|\theta) \right) \tag{5}$$

where $q$'s are arbitrary probability distributions. The free energy is limited from below by the negative log likelihood of the data $-\log p(\{h^{(t)}\})$ and this bound becomes tight when $q$ is equal to the true posterior over the hidden variables $b$ and $m$ [45]. Variational learning consists in the minimization of $F$ with respect to a constrained posterior $q$ and the model parameters $\theta$. Since for our model the number of possible configurations of $\{m_i\}$ is exponential, so we have to use a factorized form that lead us to a structured variational approximation [46]:

$$q(b^{(t)}, m^{(t)}|\theta) = \delta(\theta - \hat{\theta}) \cdot \left( q(b^{(t)}|\theta) \cdot \prod_{i=1}^{H} q(m_i^{(t)}|b^{(t)}, \theta) \right) \tag{6}$$

Substituting $P(m,b,h)$ (Eq. (2)) and $q(m,b)$ (Eq. (6)) into Eq. (5), we obtain the free energy for the occluded background model. EM alternates between minimizing $F$ with respect to the set of distributions $q(H^{(1)}), \dots q(H^{(T)})$ in the $E$ step, and minimizing $F$ with respect to $\theta$ in the $M$ step.

When updating $q(H^{(t)})$, the only constraint is $\int_{H^{(t)}} q(H^{(t)}) = 1$ and this is accounted for by using Lagrange multipliers. In summary, the pseudo-code for the EM minimization process is:

**Initialization**: choose randomly initial values for the parameters $\theta$

**E Step**: minimize $F$ w.r.t. $q$ by setting

$$q(H^{(t)}) \leftarrow P(H^{(t)}|V^{(t)}, \theta) \tag{7}$$

for each training case, given the parameters $\theta$ and the data $V^{(t)}$

**M step**: minimize $F$ w.r.t. the model parameters $\theta$ by solving

$$\frac{\partial}{\partial\theta}\log P(\theta) - \sum_t \left(\int_{H^{(t)}} q(H^{(t)}) \cdot \frac{\partial}{\partial\theta} P(H^{(t)}, V^{(t)}|\theta)\right) = 0 \tag{8}$$

This is the derivative of the expected log-probability of the complete data. For $M$ parameters, this results in a system of $M$ equations. The prior probability on the parameters (except $\psi_n$) is assumed to be uniform, i.e., $P(\theta) = \text{constant}$.

**Repeat for a fixed number of iterations or until convergence**

This process yields the following intuitive update rules:

- Cluster assignment of an image $t$ is based on the similarity of observed local measurements to what is expected in a particular background class $b$, according to the estimated parameters $\mu_b, \psi_b$, as well as the expected salient bins $\alpha_b$.

$$q(b = \tilde{b}) \propto \pi_{\tilde{b}} \cdot \prod_{i=1}^{H} \left(\frac{\alpha_{\tilde{b}i}}{q(m_i = 1|\tilde{b})} \cdot \mathcal{N}(h_i; \mu_{\tilde{b}i}, \psi_{\tilde{b}i})\right)^{q(m_i=1|\tilde{b})} \cdot \left(\frac{1-\alpha_{\tilde{b}i}}{q(m_i = 0|\tilde{b})} \cdot \mathcal{N}(h_i; \mu_n, \psi_n)\right)^{q(m_i=0|\tilde{b})} \tag{9}$$

- The masks are updated so as to balance the agreement with the overall peculiarity maps $\alpha_b$

$$\begin{aligned} q(m_i = 1|b) &\propto \alpha_b \cdot \mathcal{N}(h_i; \mu_{bi}, \psi_{bi}) \\ q(m_i = 0|b) &\propto (1-\alpha_b) \cdot \mathcal{N}(h_i; \mu_{ni}, \psi_{ni}) \\ q(m_i = *, b) &= q(m_i = *|b) \cdot q(b) \\ q(m_i = *) &= \sum_b q(m_i = *, b) \end{aligned} \tag{10}$$

where $*$ stands for 1 or 0.

- The parameters are updated to reflect the assignment statistics over all images

$$\begin{aligned} \pi_b &= \frac{1}{T} \cdot \sum_t q(b^{(t)} = b) \\ \mu_{bi} &= \frac{\sum_t q(m_i^{(t)} = 1, b^{(t)} = b) \cdot h_i^{(t)}}{\sum_t q(m_i^{(t)} = 1, b^{(t)} = b)} \\ \psi_{bi} &= \frac{\sum_t q(m_i^{(t)} = 1, b^{(t)} = b) \cdot (h_i^{(t)} - \mu_{bi})^2}{\sum_t q(m_i^{(t)} = 1, b^{(t)} = b)} \\ \mu_{ni} &= \frac{\sum_t q(m_i^{(t)} = 0) \cdot h_i^{(t)}}{\sum_t q(m_i^{(t)} = 0)} \\ \psi_{ni} &= \frac{\sum_t q(m_i^{(t)} = 0) \cdot (h_i^{(t)} - \mu_{ni})^2}{\sum_t q(m_i^{(t)} = 0)} \end{aligned} \tag{11}$$

### 3.3. Back-projection and the effect of the masks

To ease the understanding of the effects of the learning phase, we introduce here the back-projection operation. Once the learning has been performed, the saliency mask $m^{(t)}$ parameterized by the peculiarity map $\alpha_b$ represents the salient feature bins of a particular image with respect to a particular background prototype. Therefore, each image $\mathbf{z}^{(t)}$ will have $B$ saliency masks, modelled variationally by $q(m^{(t)}|b^{(t)})$ (see Eq. (6)). The back-projection depicts the effect of the $b$th peculiarity map (and thus, of the mask) directly on the images. This is given by showing on the image only those pixel values that are modelled by a salient bin; this is not an exact operation since we do not take into account the value of pixel present in the associated bin ($\mu_{b,i}$), however it gives a good approximation of what our method considers background.

For example, in Fig. 2D, three peculiarity maps are projected on a picture. The first peculiarity map is associated to a prototype shown in Fig. 2C, that models the grass, the second is associated to a prototype modeling the water while the third is associated to the road. Other projections are shown in Fig. 3; on the second column from right, it is worth noting the image with the three runners: both the street and the trees are highlighted and recognized as two different prototypes.

## 4. The proposed approach: background distillation and specialization

In our approach, the OB model is applied in a multi-scale fashion (see Fig. 4). We have $S_{max}$ levels, indexed by $s = 1 \ldots S_{max}$.

At each level, a two-step approach is performed, composed by the distillation phase (i.e., the OB model learning) and the specialization phase. At the first level, $s = 1$, all the $T$ images in their entirety are considered, by collecting in the set $I_1$ their histogram descriptors $h^{(t)}, t = 1, \ldots, T$ (Fig. 4A). The OB model is trained on $I_1$, hence generating $B$ class prototypes $\{\mu_b, \psi_b\}_{b=1,\ldots,B}$. The "optimal" value for $B$ could be selected by the user or estimated by evaluating a model selection principle, such as the maximum description length (MDL) principle [47,48], but, in this paper, model selection issues are not investigated; instead, an effective heuristics for choosing the correct $B$ is proposed later in this sections.

In the subsequent phase (Fig. 4B), namely the *BG specialization*, a similarity matrix between image descriptors $\{h^{(t)}\}$ and distilled prototypes $\{\mu_b, \psi_b\}$ is built, employing as similarity measure the intersection distance:

$$\mathcal{D}(t, b) = 1 - \frac{\sum_i \min\left(h_i^{(t)}, \mu_{bi}\right)}{h_i^{(t)}} \tag{12}$$

The intersection distance measures the percentage of the image histogram $h^{(t)}$ covered by the $b$th BG prototype.

Each $t$th image ($\mathbf{z}^{(t)}$) is then associated with the couple $\langle b_t, d_t\rangle$ where $b_t$ represents the index of the nearest prototype and $d_t$ the relative distance between the image and its nearest prototype. In formulae:

$$\langle b_t, d_t\rangle = \langle \arg\min_b D(t, b), \min_b D(t, b)\rangle \tag{13}$$

Now, all the images whose distance from their nearest prototype exceeds a threshold $\tau$ (i.e., $d_t \geqslant \tau$) are split in four non-overlapping squared sectors following a quad tree structure (Fig. 4C). All the sectors resulting from the splitting form the novel training set $\mathbf{I}_2$ (Fig. 4D). The remaining images (i.e., those for which the relation $d_t < \tau$ does hold) are labelled with $b_t$.

The intuitive idea underlying the splitting process is that if an image is not well modelled by any prototype, it could be better modelled by a potential prototype whose evidence has not

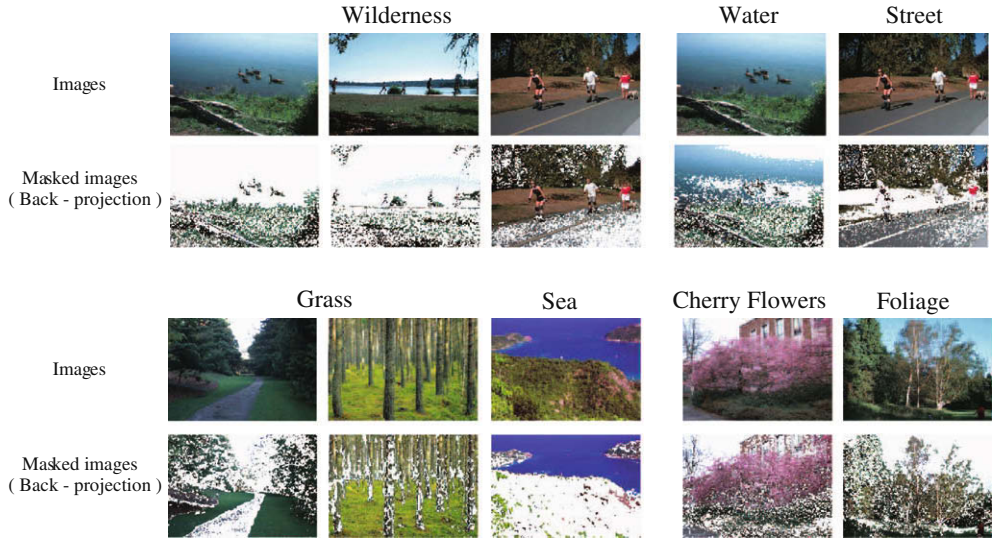


**Fig. 3.** Backprojection of the peculiarity maps. Near sample images we show a particular peculiarity map $\alpha_b$ projection. The maps refers to some prototypes found during our experiments, whose labels are written on the top of each map (see Section 5).

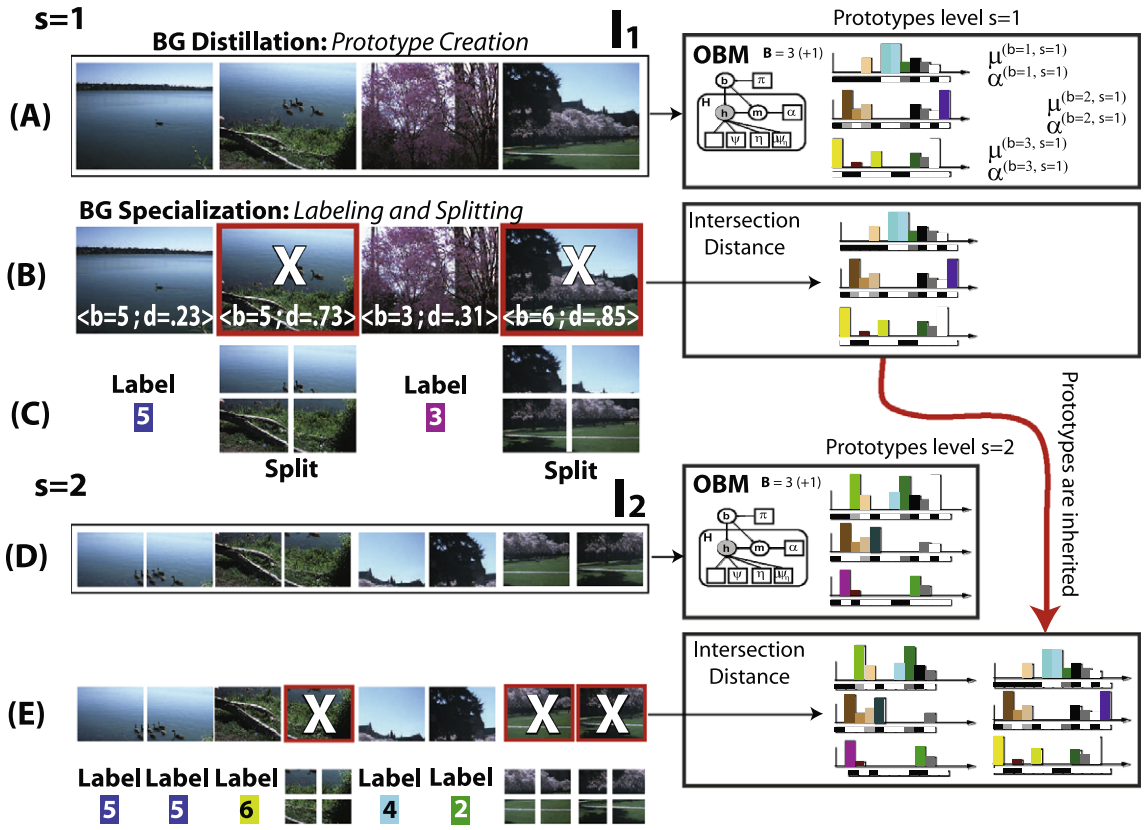


**Fig. 4.** Overview of the proposed method.

emerged in the training dataset, or that it exhibits a composite structure whose local visual components may fit with one the previously found prototypes. The threshold $\tau$ can be chosen by cross validation, although in all the experiments performed a choice of $\tau = 0.25$ leaded to good results.

This two-step process continues iteratively until a smallest sector size is considered, making sure that at each specialization step, the intersection distance between each sector and each prototype found at the upper levels is evaluated (Fig. 4E), so that the prototypes found at a given level $s$ are inherited at the lower levels.

At the end of the process, a multi-scale representation of the image is produced and stored as a quad tree. In order to create a normalized image signature, a subdivision is performed on the entire dataset. Each image is partitioned in sectors. The size of the sectors is equal to the (smallest) size of a sector at level $S_{max}$. Each sector inherits the label of the correspondent ancestor quad tree level.

In this way, each image can be described as an histogram of prototype labels, called concept occurrence vector (**COV**, See Fig. 5). We use this definition since each prototype models a natural pattern present in an image, such as sky, water, and grass.

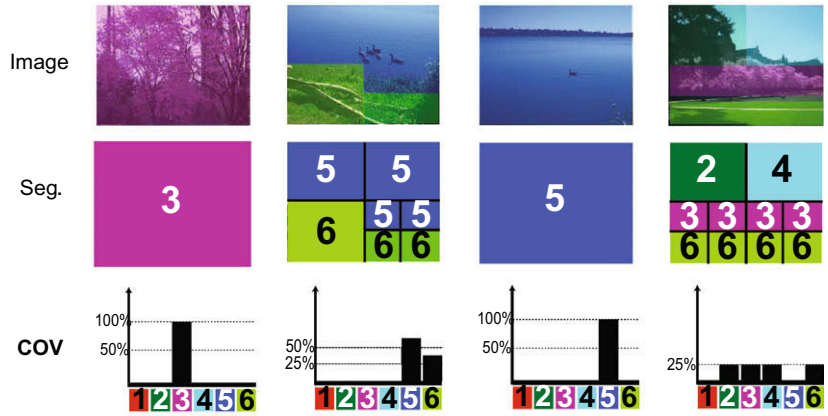


**Fig. 5.** Outcome of the representation framework: some segmentations and related *concept occurrence vector* (**COV**).

This information enables us to make a global statement about the amount of a particular concept being present in the image, e.g., "This image contains 25% grass" and, from a probabilistic point of view, represents the probability of finding a concept in that image.

This image representation allows us to model information about *which* concept appears *in which location* of an image, thus going beyond the bag-of-words paradigm, where the spatial layout of the features is lost.

At the end of the process described so far, we have at most $S_{max} \cdot B$ prototypes modeling typical, natural scene concepts with the relative peculiarity maps (see Fig. 6).

## 5. Experiments and discussion

Natural image classification is a very difficult task as boundaries between natural classes are not well defined and images across categories share much of their content. In the literature, natural scene classification often rests upon two levels of abstraction, i.e., the *concept* and the *category* levels. The concept level can be thought of as the analogue to the "latent" (topic) level in topic models [15], that, if translated into a natural image classification context, models natural constitutive elements, such as sky, rocks, sand, and snow.

Above this description level, category can be thought of as a weighted ensemble of several concepts in which the weight of each concept reflects the importance of that concept in the category definition (e.g., in the "high mountain" category, the "snow" concept has high weight). A general definition of *natural categories* has been devised accounting for different psychophysical studies [23,49–51], which witness the presence of smooth boundaries and important overlapping among categories. To test our approach, we investigate how our method works at both concept and category level, employing different kinds of natural categories.

### 5.1. Data description and experimental setup

In the experiments we consider two subsets of the Washington database [52] and one from the Corel PhotoCD database.

- *Corel Photo CD* ($\mathbf{D}_{CPC}$). We chose the same subset used in [11], probably the most complete natural image dataset. Images are divided into six categories coasts, rivers/lakes, forests, planes, mountains, sky/clouds. This categorization was introduced in [11] by combining and extending the basic-level categories introduced in [23,49].
- *Reduced Washington dataset* ($\mathbf{D}_{RW}$). We took only five categories, Arborgreens (AR), Green Lake (GL), Cherry (CH), Swiss Mountains (SM) and Greenland (GR), for a total amount of nearly 450 images. We annotated manually this dataset in two ways, in order to gather ground truth data for comparative image representation experiments. In particular, we partition all the images into non-overlapping square sectors of 1/16 of the original image size. Then, we adopt the original text labels furnished with the dataset to classify all the sectors. A more principled labelling was performed by employing the concepts definition proposed by Mojsilovic and Gomes [49]. They introduced nine local semantic concepts through the analysis of the semantic similarities and dissimilarities of a large set of images. These nine semantic concepts permit to annotate in an intuitive and precise way any landscape, forming the vocabulary **SC** = [*sky, water, grass, trunks, foliage, field, rocks, flowers, sand*]. The manual labelling turned out to be very expensive, and for this reason it has been limited to a little part of the whole dataset.
- *Natural Washington dataset* ($\mathbf{D}_{NW}$). We took all the 13 categories of natural images present in the Washington dataset: Arborgreen (AG), Australia (AU), Cannon beach (CB), Cherry (CH), Columbia George (CG), Green lake (GL) Greenland (GR), Indonesia (IN), Leafless trees (LT), San Juan (SJ), Spring flowers (SF), Swiss Mountains (SM), Yellowstone (YE), for a total amount of more than 800 images.

A selection of images is visible in Fig. 7. For what concerns the setting of the parameters, we set $s_{MAX} = 3$,[4] thus managing sectors large 1/16th of the original image size. At each background distillation step, we learn $B = 6$ prototypes, considering as *valid* only those prototypes which were assigned to more than five images/sectors after the BG specialization step. This heuristic model selection strategy worked well but a more principled procedure is currently under study.

### 5.2. Validation of the prototypes

The first test consists in applying our approach to the $\mathbf{D}_{RW}$ database. At the end of the whole multi-scale process, we identified eight prototypes for $\mathbf{D}_{RW}$, estimated at different levels (three at the first level and five at the second one). Image segmentations through the sector labelling are visible in Fig. 8. Labels were given by following the strategy explained later in this section.

As test for the significance of the obtained prototypes, we consider the ground truth text annotations of the reduced Washington

[4] This parameter depends on the resolution of the images.

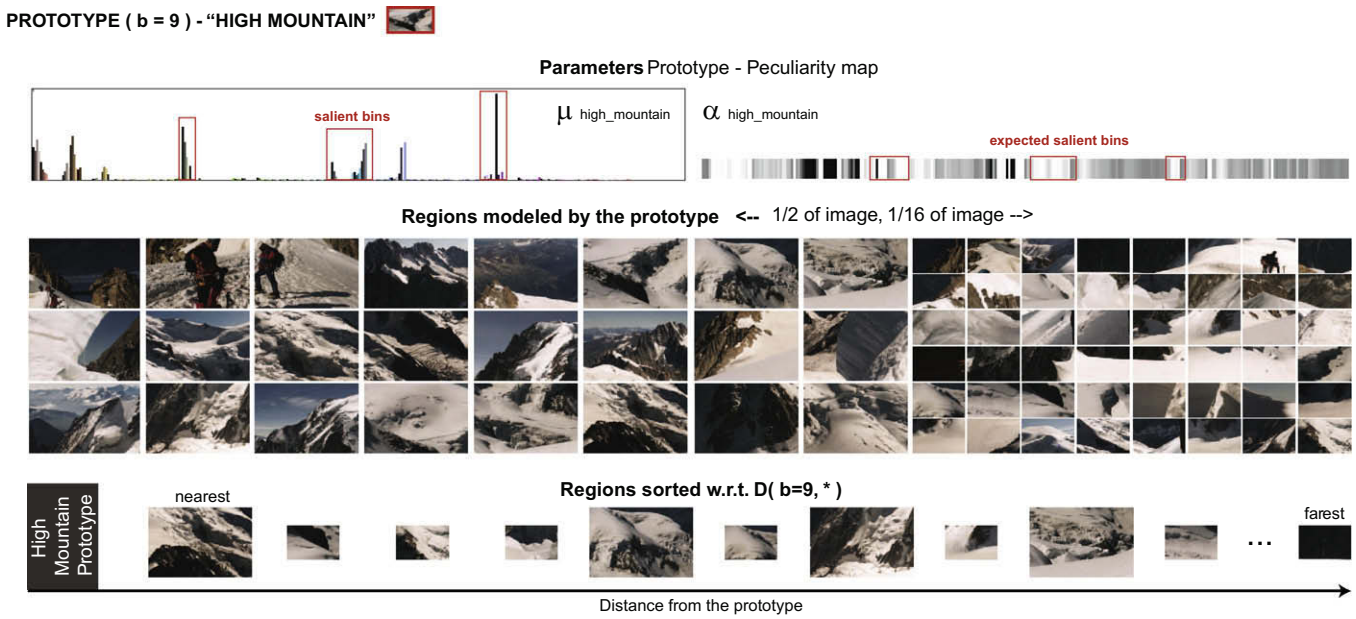


**Fig. 6.** Class (named) "High mountains". The prototype $\mu$ and its peculiarity map $\alpha$. In the second row, examples of sectors labelled with that prototype. On the bottom, the sectors ordered by the distance to the prototype.

database ($\mathbf{D}_{RW}$), and we draw a correspondence table of that annotations with our prototypes. In such table, we report those correspondences that hold more than the 90% of the examined cases (see Table 1). In practice, we list those correspondences that linked at least 90 occurrences of the same textual annotation with 100 sectors labelled with the same BG prototype. The typicality of the label assignment performed by our approach is evident, in the sense that each prototype refers to particular natural concept. In order to summarize all the valid text correspondences in a unique concept label, we assign to the eight ordered prototypes the following text labels, forming the concept vocabulary $\mathbf{SC}_{rw}$ = [*wilderness, cherry flowers,water, field/bushes, foliage, street, high mountain, cherry gems*].

In Fig. 9a, for each prototype its nearest image/sector is shown.

The second test considers the labelling performed by using the concepts of **SC**. Here, after learning we have learned the prototypes $\mathbf{SC}_{cpc}$ and $\mathbf{SC}_{rw}$ (see Fig. 9), respectively, for we calculate the number of times that each of the prototypes is assigned by the proposed method to a region labelled by each of the concept present in **SC**, forming thus a similarity matrix depicted in Fig. 10a.[5] This result confirms what was found in the previous test, i.e., that our approach is able to naturally segregate natural patterns, disregarding clutter elements, providing prototypes which can be considered as accurate natural concepts, in the sense defined by [15].

In the same way, we apply our approach to the Corel dataset, whose per-sector ground-truth labelling is given in [11]. We obtain 12 prototypes (five at the first level, seven at the second level). In Fig. 9b the list of the nearest image/sector to each prototype is reported. The similarity matrix that explains the correspondence between our prototypes and their natural concept is given in Fig. 10b. To ease the understanding of the matrix, we manually assign to each prototype a text label explaining the most prominent intuitive natural concept forming $\mathbf{SC}_{cpc}$. Even in this case, we can observe that the similarity matrix is peaked, highlighting the capability of our prototype to capture well defined natural concepts. Moreover, it is worth to note that: (1) our method avoids to produce wrong similarities (i.e., there is no similarity between rocks and sea) and (2) the approach uses simple image descriptors such as the color histogram, which have a very limited expressiveness. Working with more complex descriptors in a bag-of-words framework could lead to a more powerful and precise prototype definition.

### 5.3. Image classification

In order to test the ability of our image concept representation to build category definitions, in the sense of ensemble of natural concepts of [15], we adopt the category partitions proposed in the Washington database (provided by the database authors) for what concerns $\mathbf{D}_{NW}$ (we drop the restricted $\mathbf{D}_{RW}$ database), and the categories introduced in [11] for what concerns the $\mathbf{D}_{CPC}$ database. After having learnt the concept representations for each database, producing the vocabulary $\mathbf{SC}_{nw}$ and $\mathbf{SC}_{cpc}$, we calculate the concept occurrence vector **COV** for each image (see Section 4), and subsequently we calculate $p^c$, the "category **COV**", built by calculating the mean over all the **COV** of the images of a given category.

The resulting $p^c$ are depicted in Fig. 11.

The classification policy is the following: given an image and its **COV**, we assign it to the nearest category via L2-norm between **COV** and $p^c$; the resulting procedure is unsupervised for what concerns the concepts but supervised at category level.

We compared our method with the discriminative method of [11], employing the same classification policy, and with the generative topic model presented in [5].

In [11], nine support vector machines are learnt in order to discriminate between the nine natural semantic concepts **SC** found in [49]. Each image is then partitioned in 100 non-overlapping fixed-regions and each region is assigned to a concept via SVM classification.

In [5], Latent Dirichlet Allocation (LDA) is used to classify scenes. In particular, in this work a visible class variable is added to the original LDA graphical model inferring, separately for each class, a prior $\alpha_{class}$ over the theme distributions of the single images.

[5] We used the ground-truth labels provided by [11] for $\mathbf{D}_{CPC}$ and we repeated the labelling procedure of [11] to create the ground-truth for $\mathbf{D}_{RW}$.

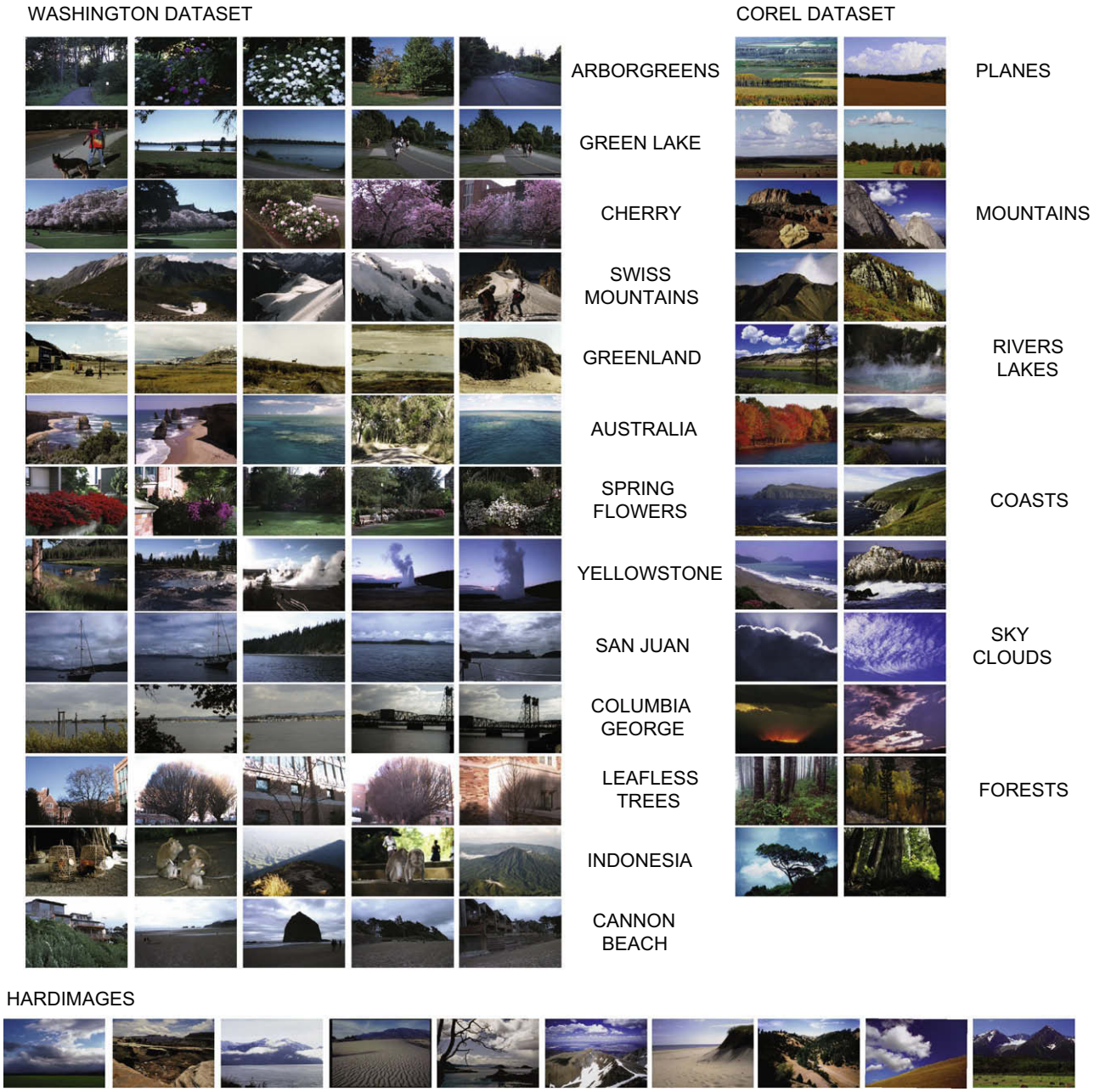


**Fig. 7.** Some examples of the images used for each category. Note how in many images the presence of foreground is evident, i.e., visual object(s) not characteristic of any natural category. On the bottom we report some images that would be difficult to classify even for humans.

The classification is performed in a generative way, i.e., calculating the likelihood of unseen images under each class parameters.

It is worth to note that our method presents similarities with [11] for what concern image description but it has three important conceptual differences: (1) the concepts are not a-priori defined but extracted from the dataset, (2) the method does not require manual preprocessing operations, and (3) the images are subdivided only if necessary.

We performed image classification on each dataset mentioned. Confusion matrices of results of our method are shown in Fig. 12, while overall classification rates are reported in Table 2, where we report the accuracy for the proposed method for 2 choice of $S_{max}$. The column "[11] *with OB*)" refers to the method in [11] previously presented, with the SVM learned using the concepts estimated by our method (see Fig. 9). In practice, for each concept $b$ found by our method, we sort the image sectors according to their distance from the $b$th prototype (see Fig. 6), and we learn an SVM with RBF kernel with the 15 nearest images. In formulae the training set used is:

$$\text{TRAINING}_b = \{h_b^{(t)} \mid D_{sort}(t,b) \text{ is ranked between 1 and 15}\} \quad (14)$$

The optimal SVM parameters are found via cross-validation. As additional test for evaluating the contribute of the generative modeling, we repeated the classification task using a simple mixture of Gaussians (MoG) instead of OB model. It is worth to note that mixture of Gaussians is a particular case of OB model with the masks $m_i^{(t)} = 1 \ \forall i, \ t$, i.e., considering salient each pixel. The usage of mixture of Gaussians degrades the performance of 7-10 % for these datasets (See Table 2).

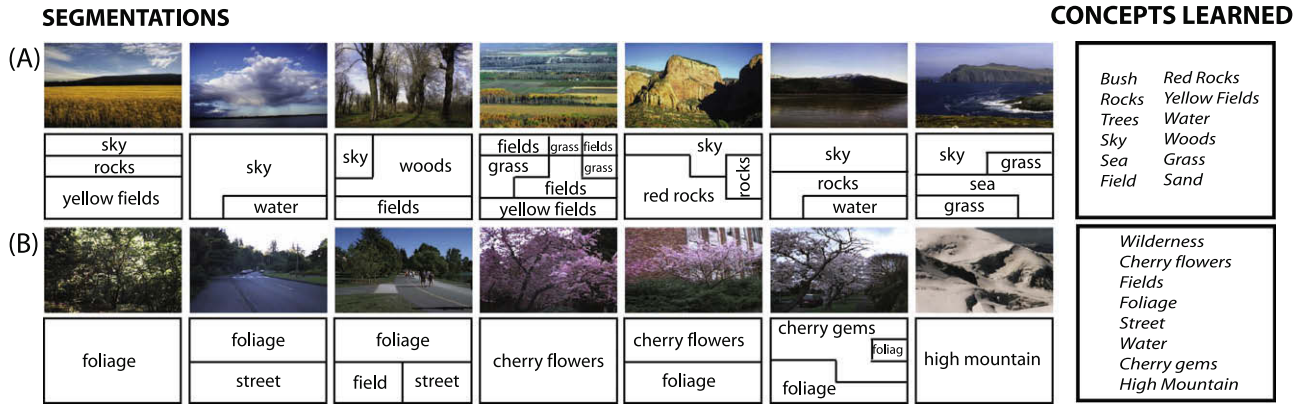


**Fig. 8.** Some examples of the segmentations obtained by our approach (A) Corel dataset and (B) Washington dataset. On the right, the concepts learnt for both datasets.

**Table 1**
Correspondence between concepts. The text annotations for each prototype are ordered in decreasing order with respect to relative frequency.

| OB model prototypes | Washington annotations |
|---|---|
| Prototype 1 | Ground, grass, mountain |
| Prototype 2 | Tree, trunk |
| Prototype 3 | Water |
| Prototype 4 | Grass, bush |
| Prototype 5 | Bushes, fern, lily |
| Prototype 6 | Street, trail, sidewalk |
| Prototype 7 | Rocks, snow, ice, clear sky |
| Prototype 8 | Flowers, cherry tree |

The results are easy to understand; OB model *per se* reaches good performances in all dataset providing similar performances to [11,5]. Moreover, combining the concepts of our approach with the discriminative and finer method of [11], we obtain the best results on all the datasets.

## 6. Conclusions

In this paper, a generative model for multi-scale image representations from an image dataset is presented. The final image representation is functional to tasks like natural image classification and natural scene categorization. The method is based on a generative image model able to distinguish, at different levels, the background information from the foreground elements in whatever position, thus allowing to focus on the scene information only. In summary, given an image database, the designed generative model allows to disregard the entities which are not in accordance with the background data, and this process is iterated at different levels down to a certain (a priori fixed) sector size. The image (sector) subdivision is carried out automatically in case its content is not homogeneous enough, hence requiring a further split to better identify the constitutive elements of the scene considered. This method differs from the approaches in the literature in several aspects. Although there are other generative methods for image classification, none of them attempts to discriminate actual useful information given by the background from other clutter information present in the image and this much improves the classification accuracy.

The experiments show how the proposed method is able to select salient concepts from an image dataset of natural scenes of different types. Concepts are selected in a robust way by extracting the so-called saliency masks which make possible to work also with personal pictures typically affected by clutter, i.e., information not useful for natural scene classification (e.g., persons, faces,

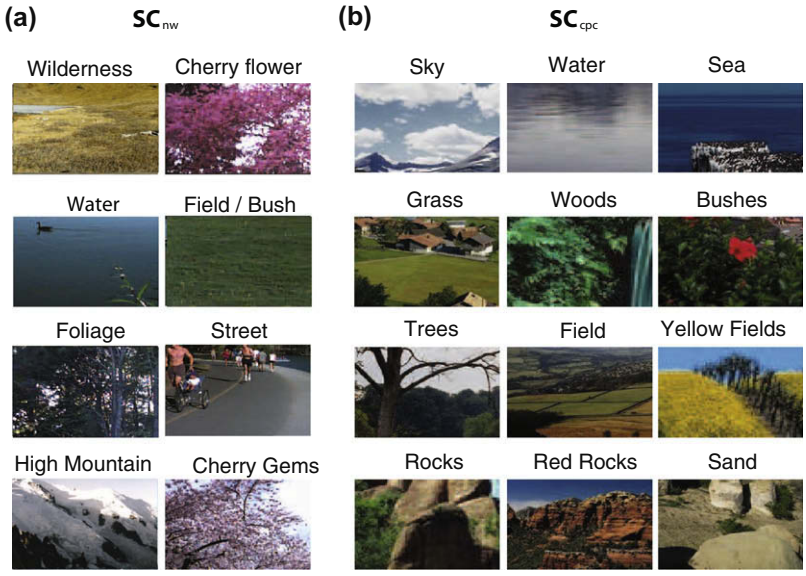


**Fig. 9.** Concept listings: for each $\mathbf{D}_{RW}$ (a) and $\mathbf{D}_{CPC}$ (b), we show the nearest image (or sector) to each prototype.

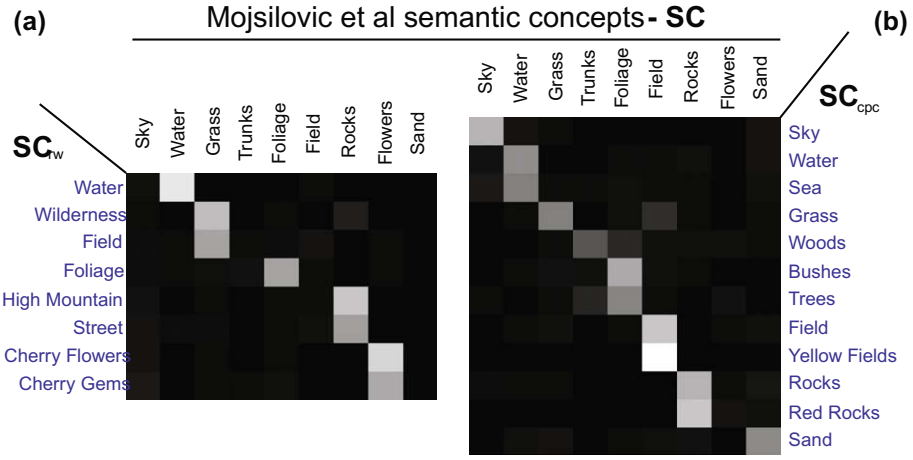


**Fig. 10.** Concept listings: similarity matrices between the prototypes found by the occluded background model ($\mathbf{SC}_{rw}$ and $\mathbf{SC}_{cpc}$) and the semantic concepts **SC** introduced in [49].

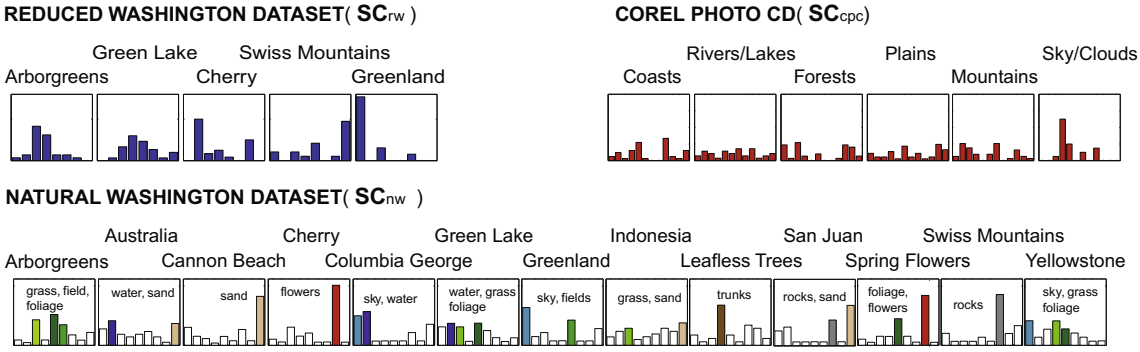


**Fig. 11.** Category concept occurrence vectors ($p^c$) for all the dataset analyzed. Each bin is proportional to the probability of occurrence of a particular concept. For $\mathbf{D}_{NW}$ we highlight the most frequent prototype concept for each category.

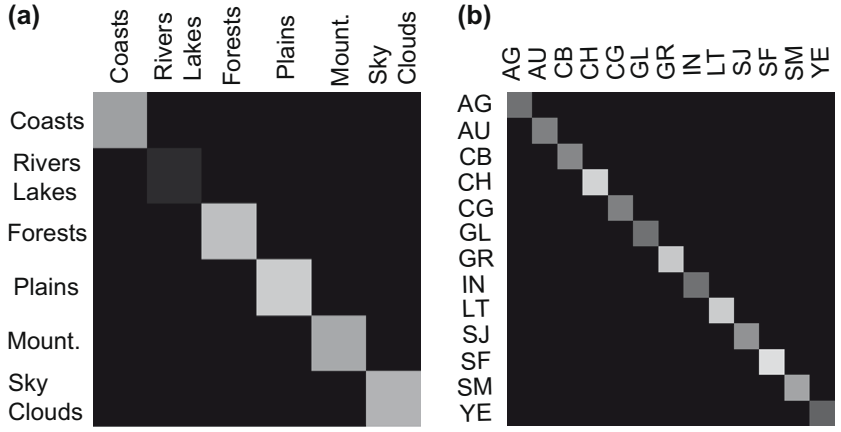


**Fig. 12.** The confusion matrices for the OB model classification show the major sources of ambiguity. (A) Natural Washington dataset and (B) Corel dataset. Numerical results are reported in Table 2

**Table 2**
Classification results. In bold our results.

| Dataset | Chance | OB Model $S_{max}$ = 3 | OB Model $S_{max}$ = 2 | MoG |
|---|---|---|---|---|
| $\mathbf{D}_{CPC}$ | 16.7% | **54.9%** | **41.2%** | 47.5% |
| $\mathbf{D}_{NW}$ | 7.7% | **57.6%** | **49.2%** | 48.0% |
| | | [11] | [11] with OB model concepts | [5] |
| $\mathbf{D}_{CPC}$ | | 71.7% | **80.1%** | 52.5% |
| $\mathbf{D}_{NW}$ | | 55.1% | **61.3%** | 53.5% |

etc.). Further efforts are planned to analyze in more in detail the performances of our method on the retrieval task, by testing it on larger standard databases.

## References

[1] J. Henderson, Introduction to real-world scene perception, Visual Cognition 12 (3) (2005) 849–851.
[2] G. Heitz, D. Koller, Learning spatial context: using stuff to find things, in: Proceedings of the European Conference on Computer Vision (ECCV 2008), 2008, pp. 30–43.

[3] A. Oliva, A. Torralba, Building the gist of a scene: the role of global image features in recognition, Progress in Brain Research: Visual Perception 155 (2006) 23–36.
[4] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, Journal of machine Learning Research 3 (2003) 993–1022.
[5] F. Li, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, 2005, pp. 524–531.
[6] J. Vogel, Semantic Scene Modeling and Retrieval, Selected Readings in Vision and Graphics, vol. 33, Hartung-Gorre Verlag, Konstanz, 2004.
[7] J. Vogel, A. Schwaninger, C. Wallraven, H.H. Bulthoff, Categorization of natural scenes: local vs. global information, in: Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization, 2006, pp. 33-40.
[8] C. Stauffer, W. Grimson, Adaptive background mixture models for real-time tracking, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 1999), vol. 2, 1999, pp. 246–252.
[9] A. Torralba, A. Oliva, Statistics of natural image categories, Network: Computation in Neural Systems 14 (3) (2003) 391–412.
[10] J. Vogel, B. Schiele, A semantic typicality measure for natural scene categorization., in: DAGM-Symposium, 2004, pp. 195–203.
[11] J. Vogel, B. Schiele, Semantic modeling of natural scenes for content-based image retrieval, International Journal of Computer Vision 72 (2) (2007) 133–157.
[12] D.D. Lewis, Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval, Springer-Verlag, 1998. pp. 4–15.
[13] C. Dance, J. Willamowski, L. Fan, C. Bray, G. Csurka, Visual categorization with bags of keypoints, in: Proceedings of the ECCV International Workshop on Statistical Learning in Computer Vision, 2004, pp. 1–22.
[14] J. Sivic, B. Russell, A. Efros, A. Zisserman, B. Freeman, Discovering objects and their location in images, in: Proceedings of the International Conference on Computer Vision (ICCV 2005), 2005, pp. 370–377.
[15] A. Bosch, A. Zisserman, X. Munoz, Scene classification via plsa, in: Proceedings of the European Conference on Computer Vision (ECCV 2006), 2006, pp. 517–530.
[16] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D.M. Blei, M.I. Jordan, Matching words and pictures, Journal of Machine Learning Research 3 (2003) 1107–1135.
[17] G. Salton, M.J. Mcgill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, USA, 1986.
[18] T. Hofmann, Probabilistic latent semantic indexing, in: SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, New York, NY, USA, 1999, pp. 50–57.
[19] S. Agarwal, D. Roth, Learning a sparse representation for object detection, in: Proceedings of the European Conference on Computer Vision (ECCV 2002), 2002, pp. 113–130.
[20] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, L.V. Gool, Modeling scenes with local descriptors and latent aspects, in: Proceedings of the International Conference on Computer Vision (ICCV 2005), 2005, pp. 883–890.
[21] D. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the International Conference on Computer Vision (ICCV 1999), 1999, pp. 1150–1157.
[22] R. Duda, P. Hart, D. Stork, Pattern Classification, John Wiley and Sons, 2001.
[23] B. Tversky, K. Hemenway, Categories of environmental scenes, Cognitive Psychology (15) (1983) 121–149.
[24] A. Vailaya, M. Figueiredo, A. Jain, H.-J. Zhang, Image classification for content-based indexing, IEEE Transactions on Image Processing 10 (1) (2001) 117–130.
[25] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, A thousand words in a scene, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (9) (2007) 1575–1589.
[26] J. van Gemert, J. Geusebroek, C. Veenman, C. Snoek, A. Smeulders, Robust scene categorization by learning image statistics in context, in: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop (CVPRW 2006), 2006, pp. 105–112.
[27] F. Perronnin, Universal and adapted vocabularies for generic visual categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (7) (2008) 1243–1256.
[28] J. Geusebroek, A. Smeulders, A six-stimulus theory for stochastic texture, International Journal of Computer Vision 62 (1–2) (2005) 7–16.
[29] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, in: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop (CVPRW 2006), 2006, pp. 13–20.
[30] R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, Learning object categories from googles image search, in: Proceedings of the International Conference on Computer Vision (ICCV 2005), 2005, pp, 1816–1823.
[31] D. Liu, D. Chen, T. Chen, Latent layout analysis for discovering objects in images, in: Proceedings of the International Conference on Pattern Recognition (ICPR 2006), 2006, pp. 468–471.
[32] D. Liu, T. Chen, Semantic-shift for unsupervised object detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop (CVPRW 2006), 2006, p. 16.
[33] D. Liu, T. Chen, Unsupervised image categorization and object localization using topic models and correspondences between images, in: Proceedings of the International Conference on Computer Vision (ICCV 2007), 2007, pp. 1–7.
[34] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR 2006), 2006, pp. 2169–2178.
[35] E. Hadjidemetriou, M. Grossberg, S. Nayar, Multiresolution histograms and their use for recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (7) (2004) 831–847.
[36] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12) (2000) 1349–1380.
[37] R. Veltkamp, M. Tanase, D. Sent, Features in content-based image retrieval systems: a survey, in: State-of-the-Art in Content-Based Image and Video Retrieval, Kluwer, 1999, pp. 97–124.
[38] Y. Rui, T. Huang, S. Chang, Image retrieval: current techniques, promising directions and open issues, Journal of Visual Communication and Image Representation 10 (4) (1999) 39–62.
[39] B. Russell, W. Freeman, A. Efros, J. Sivic, A. Zisserman, Using multiple segmentations to discover objects and their extent in image collections, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2006), 2006, pp. 1605–1614.
[40] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 888–905.
[41] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, International Journal of Computer Vision 42 (3) (2001) 145–175.
[42] B. Frey, N. Jojic, A comparison of algorithms for inference and learning in probabilistic graphical models, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (9) (2005) 1392–1413.
[43] N. Jojic, B. Frey, Learning flexible sprites in video layers, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR 2001), 2001, pp. 199–206.
[44] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society Series B 39 (1977) 1–38.
[45] M. Jordan, Z. Ghahramani, T. Jaakkola, L. Saul, An introduction to variational methods for graphical models, Machine Learning 37 (2) (1999) 183–233.
[46] Z. Ghahramani, On structured variational approximations, Tech. Rep. CRG-TR-97-1, 1997.
[47] M. Figueiredo, J. Leitao, A. Jain, On fitting mixture models, in: E. Hancock, M. Pellilo (Eds.), Energy Minimization Methods in Computer Vision and Pattern Recognition, Springer-Verlag, 1999, pp. 54–69.
[48] J. Rissanen, Stochastic complexity and modeling, The Annals of Statistics 14 (3) (1986) 1080–1100.
[49] R.B. Mojsilovic, A.J. Gomes, Semantic-friendly indexing and querying of images based on the extraction of the objective semantic cues, International Journal of Computer Vision 42 (2004) 79–107.
[50] E. Rosch, Principles of Categorization in Cognition and Categorization, John Wiley & Sons Inc, 1978. pp. 27–48.
[51] E. Rosch, C.B. Mervis, Family resemblances: studies in the internal structure of categories, Cognitive Psychology 7 (4) (1975) 573–605.
[52] Washington Ground Truth Image Database, 2004. <http://www.cs.washington.edu/research/imagedatabase/>.