Unsupervised Learning of Saliency Concepts for Natural Image Classification and Retrieval

A. Perina, M. Cristani, and V. Murino

Dipartimento di Informatica, Università degli studi di Verona
Strada le Grazie 15, 37134 Verona, Italia
{alessandro.perina, marco.cristani, vittorio.murino}@univr.it

Abstract. In this paper, a novel multi-scale, statistical approach for natural image representation is presented. The approach selects, at different scales, sets of features that represent exclusively the most typical visual elements of several natural scene categories, disregarding other non-characteristic, clutter, elements. Such features provide also a robust image visual signature, useful for scene understanding, image classification and retrieval. The approach lies upon a structured generative model efficiently trained through variational learning. Results regarding image classification and retrieval prove the goodness of the approach.

Keywords: Image analysis, Image classification, Unsupervised learning.

1 Introduction

In the machine learning literature, a used definition of the term "scene" is the one of *a semantically coherent, namable human-scaled view of a real world environment* [1]. Classifying accurately the category of an imaged natural scene is highly useful in a wide variety of tasks: other than the mere classification of the environment, scene categorization helps also in object recognition by providing a conditional constraint on the acceptable semantic labels of the objects identities (e.g., it is rare to see a shark in a mountain environment). In the literature, all the image categorization methods can be explained as composed by two main phases: 1) representation and 2) classification.

In the first phase, features are extracted from a training image dataset, in a local or global way, and then are opportunely organized in clusters, each one representing an image category. In the second phase, actually implementing the content-based image retrieval, a query image is classified as belonging to one of the clusters. It is easy to be convinced that the representation phase plays a key role for general image categorization, and this is especially true for natural scene categorization. In this case, the representation phase should extract features belonging strictly to the natural scene depicted in the image, while ignoring all the *uncharacteristic objects* present in it in order to facilitate and make effective the classification task. With the term "uncharacteristic object" we refer here to whatever item which occurs rarely with the same appearance in a natural scene: face close-ups, entire bodies, out-of-context objects are normally *surrounded* by

J. Ruiz-Shulcloper and W.G. Kropatsch (Eds.): CIARP 2008, LNCS 5197, pp. 169–177, 2008.

natural scenes, but they are not characteristic of such scenes. For the sake of convenience, we term as background (**BG**) the scene we want to categorize, and with foreground (**FG**) every uncharacteristic object which does not belong to it.

In this paper, we propose a statistical method for the representation and classification of *generic* natural scenes, e.g., photos. The proposed method deals with generic natural scenes as it is able to automatically discriminate between the **BG** scene and **FG** objects by building a dictionary of **BG** visual patterns useful for representation and classification. The method is unsupervised, in the sense that no hand-labelling is needed for the formation of the dictionary. It is based on a structured generative graphical model whose only parameters to be set by the user are the number of different classes of the expected scenes to be considered. Variational learning is employed to keep the computational cost of the model training in acceptable limits. By providing generic visual patterns as input, the proposed method permits to *understand* the typical appearance of a natural scene, other than codifying it for further classification and retrieval. Finally, our method is multi-level, i.e., the images are inspected at different levels of detail, exploiting a quad-tree based image analysis strategy.

As a matter of fact, several methods for scene categorization work with images in which no **FG** objects are present, like [2,3]. The task is in this case facilitated with respect to the problem addressed in this paper: typically, personal photos (from vacations, for instance) may contain friends or persons, together with the natural environment, so our approach can also deal with the categorization of general "personal" pictures, unlike the methods present in the literature [3].

The rest of the paper is organized as follows. In Section 2, the generative model at the basis of the discrimination between the **BG** scene and the **FG** objects is described. Section 3 details the multi-level approach, called distillation-specialization, able to classify and cluster the database of images. In Section 4, results on classification and retrieval using different data sets are reported, also providing comparative figure of merits. Moreover, still in this Section, the obtained results are commented and final considerations are drawn.

2 The Occluded Background Model

In this study, we propose a generative model called *OccludedBackground* (OB) model which is essentially a mixture model (b is the mixture variable) with a feature selection variable m which aims at highlighting the salient features of a particular class (i.e., cluster).

The model is applied to a generic robust image description: the quantized color histogram in HSV space $\mathbf{h} = h_1, \ldots, h_i, \ldots h_H$, where H is the number of bins. Each bin value is modeled by a Gaussian function with parameters μ and ψ. The core idea is to extract from a pool of color histograms a set of B **BG** histograms prototypes $\mu^{(b)}$ (the cluster centroids), and a set of B related *saliency* **BG** masks $\alpha^{(b)}$. Each b-th **BG** mask selects, weighting properly, the color bins which are more representative for the b-th cluster.

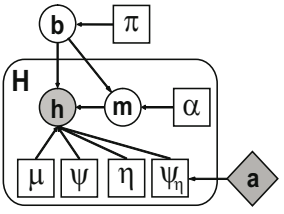

Fig. 1. Occluded background Bayesian network model. Circles represent the hidden variables, shaded circles are the visible variables, squares are the parameters and diamonds represent the constants.

Using the standard notation for Bayesian networks, the OB model is shown in Fig.1. The key issue here consists in conditionally linking the mask random variable **m** to the **BG** class and keeping a dump class which models all the **FG** objects so that an histogram bin h_i is generated starting from a prototype $\mu_i^{(b)}$ if that bin is salient, otherwise from the dump class.

In this way, the generative process that forms an observed image histogram is the following one:

1. Choose a prototype class $b \sim p(b|\pi)$ for each image, where $b = \{1, \ldots, B\}$. B is the total number of background prototypes at the current level. π is a B-dimensional array of a multinomial vector representing the prior over the **BG** classes.
2. For a particular image histogram generated by the b-th prototype, we want to determine which bins are significant for that prototype. This is done by choosing the image binary mask $m_i \sim p(m_i|\alpha_i^{(b)})$, with $m_i \in \{0, 1\}$, where $m_i = 1$ indicates that the i-th bin of the image is a salient background bin, so contributing to form the background class prototype $\mu_i^{(b)}$ and the related saliency mask bin $\alpha_i^{(b)}$. The saliency mask $\alpha^{(b)}$ represents the prior on the values of the masks **m** given the class b, and may assume values in the interval $[0, 1]$.
3. Finally, the histogram bins are selected independently, given the mask and the prototype class, extracting them from a Gaussian distribution of parameters $\mu^{(b)}$ and $\psi^{(b)}$ for the salient bins, and from a high variance Gaussian distribution of parameters η and ψ_η, for the non-salient bins (modeling the dump class).

This generative process is modeled by the following joint distribution:

$$P(h, m, b) = \pi_b \cdot \prod_{i=1}^{H} (\alpha_i^{(b)})^{m_i} (1 - \alpha_i^{(b)})^{1-m_i} \cdot \mathcal{N}(h_i; \mu_i^{(b)}, \psi_i^{(b)})^{m_i} \cdot \mathcal{N}(h_i; \eta; \psi_\eta)^{1-m_i}$$

were $\mathcal{N}(h; \mu, \psi)$ is the Gaussian density function on h with mean μ and variance ψ. If needed a gamma prior $\Gamma(\psi_\eta; \mathbf{a})$ on ψ_η with parameter **a** can be used in order to keep ψ_η high (i.e., 60% - 100% of the dynamic range of the training data).

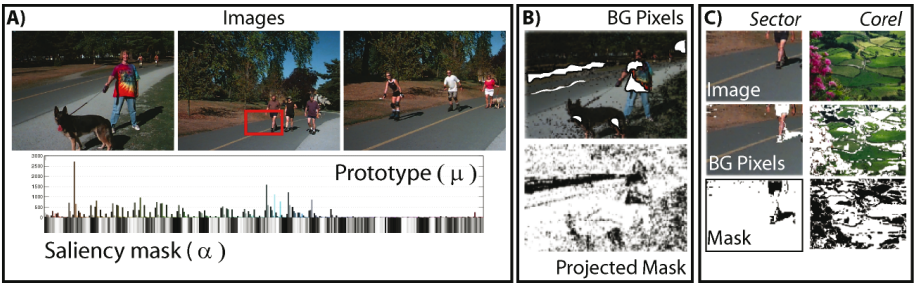

Fig. 2. A) Three images modeled by the histogram prototype shown below them. B) The BG pixels and the relative binary mask. C) Two images, their background pixel and the related binary mask.

The effect of the saliency mask is visually shown in Fig. 2-B and C. Figure 2-B shows which pixels of an image are taken into account to calculate to prototype shown in Fig. 2A. These kind of images can be obtained by retro-projection, that is drawing the pixel intensities if they are modeled by the salient bins of the histogram. Notice how the red shirt of the boy standing with his dog is considered as **FG** by the saliency mask.

2.1 Learning: Variational Inference

The model learning consists in fitting reasonable values for the (hidden) variables b and m, so as to find good (hidden) parameters estimates for all model variables. We learn the model following a Maximum Likelihood approach through Generalized Expectation Maximization (GEM)[4]. Exact inference using Expectation Maximization is unfeasible, being the computation of the posterior over the hidden variables exponential in H. The solution is to approximate the true posterior distribution $P(m, b|h)$ with a simpler generic distribution $Q(\{m, b\})$, defined assuming the training samples independent and identically distributed. For the OB model, this yield to $Q(\mathbf{h}) = \prod_{n=1}^{N} q(m^{(n)}) \cdot q(b^{(n)})$ where $m^(n)$ is the mask variable, $b^(n)$ is the **BG** prototype index variable related to the n-th histogram, respectively, $q(\cdot)$ represents a multinomial distribution, and N is the total number of samples present in a dataset. The formulation used is known as *mean field form* and assumes the independence of the hidden variables given the data.

For each observed histogram, the approximated (log) posterior over the hidden variables is:

$$\log P(b, m|h) = \sum_{n=1}^{N} \log \left(q(b^{(n)}) \prod_i q(m_i^{(n)}) \right) \tag{1}$$

GEM maximizes the bound on the (log) probability of the data:

$$logP(\mathbf{h}) \geq \sum_{n=1}^{N} \sum_b \sum_m \left[q(b) \prod_i q(m_i) \cdot \log \frac{P(h, b, m_i)}{q(b^{(n)}) \prod_i q(m_i^{(n)})} \right] \tag{2}$$

The update rules are thus obtained calculating the derivatives of the right side term of Eq.2 with respect to each hidden variable (E-Step) and each parameter $\theta = \{\mu, \psi, \eta, \psi_\eta, \alpha, \pi\}$ (M-Step).

3 The Proposed Approach: Distillation - Specialization

The OB model, which exploits the **FG**/**BG** feature selection and modelling, is applied in a multi-scale fashion. We have S scales, indexed by $s = 1 \dots S_{max}$. At each scale, a two-step approach is performed, composed by the Distillation phase (i.e., the OB model learning) and the Specialization phase. At the first scale, $s = 1$, all the N *whole* images are considered, by collecting in the set I_1 their histogram descriptors $h^{(n)}, n = 1, \dots, N$ (in our case, HSV histograms). The OB model is trained on I_1, generating B class prototypes $\mu^{(b)}$ (Fig.3-A). In the subsequent phase, namely the *BG specialization*, a similarity matrix among image descriptors $\{h^{(n)}\}$ and distilled prototypes $\{\mu^{(b)}\}$ is built, employing as similarity measure the intersection distance:

$$\mathcal{D}(n, b) = 1 - \frac{\sum_{bin} \min\left(h^{(n)}, \mu^{(b)}\right)}{|h^{(n)}|} \tag{3}$$

The intersection distance measures the percentage of the image histogram $h^{(n)}$ covered by the *b-th* **BG** prototype. Each n-th image is then labeled with the couple $< \mathbf{b}_n, \mathbf{d}_n >$ where $\mathbf{b}_n$ represents the index of the nearest prototype, $\mathbf{b}_n = \arg\min_b D(n, b)$, and $\mathbf{d}_n$ the relative distance between the image and its nearest prototype $\mathbf{d}_n = \min_b D(n, b)$. (Fig.3-B). Now, every images whose distance from their nearest prototype exceeds a threshold τ (i.e. $\mathbf{d}_n \geq \tau$) are split in four non-overlapping regions, and these regions are merged to form the set $\mathbf{I}_2$ (Fig.3-C and D). All the other images (i.e. $\mathbf{d}_n < \tau$) are labeled with the nearest prototype label $\mathbf{b}_n$ (Fig.3-C).

The intuitive justification is that if an image is not well modeled by any prototype, it means that the image can be better modeled by another prototype not yet found or it has a composite structure not explainable by considering the image as a whole. The threshold τ can be chosen by cross validation, although it is intuitive that a choice of $\tau = 0.5$ leads to good results.

The two-step process continues iteratively until a smallest region size is considered, paying attention that at each specialization step, the intersection distance between each image (or part of it) and each prototype found at the superior levels (Fig.3-E) is evaluated, so that the prototypes found at a given level s are inherited at the lower levels.

At the end of the process, a hierarchical representation of the image is produced, in which an image is subdivided in a quad-tree fashion. In order to create a normalized image signature, a subdivision is performed on the entire dataset. Each image is partitioned in regions. The size of the regions is equal to the (smallest) size of a region at level S_{max}. Each region inherits the label of the quad-tree sector over which it lies. In this way, each image is described as an histogram of concepts, called Concept Occurrence Vector (**COV**). This representation serves

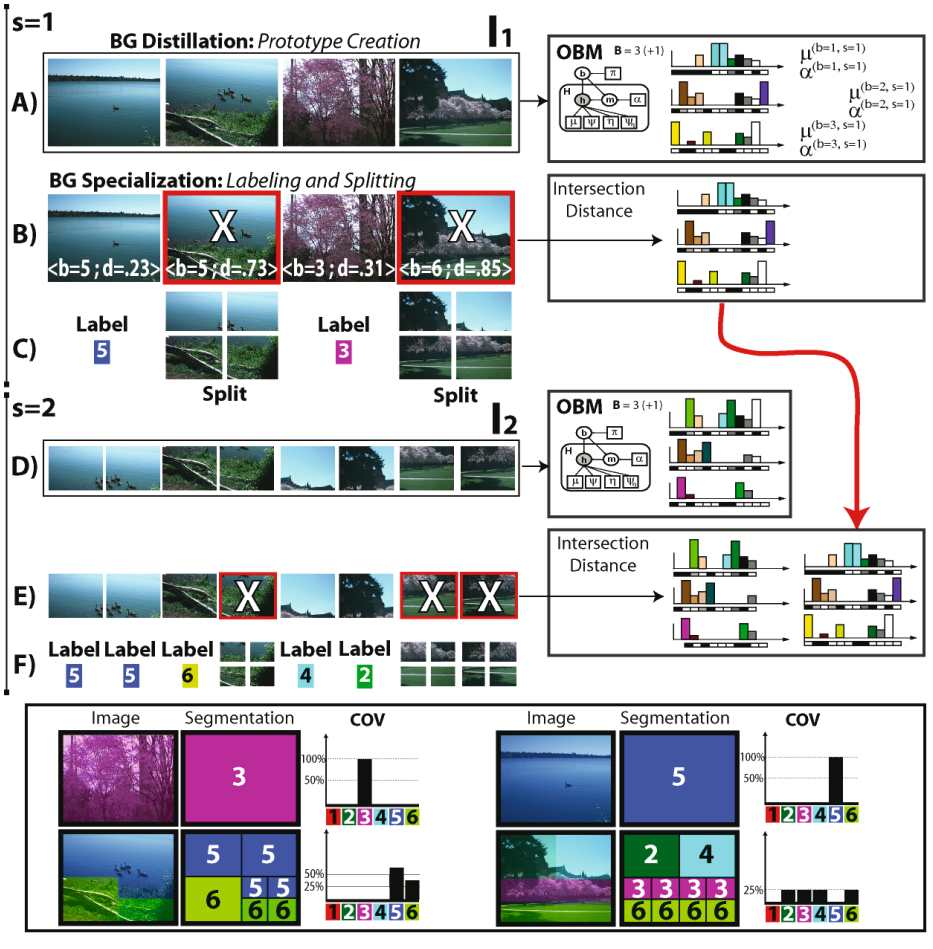

Fig. 3. Overview of the proposed method: A) The *BG* Distillation aims at creating the prototypes representing occurring semantical concepts of the dataset, B) the **BG** Specialization tries to assign a concept to each image (or part of it) B-C). If an image is too far from its nearest prototype it will be split and the whole process is repeated until the necessary level of detail is reached. On the bottom of the figure, the segmentations and the related *concept occurrence vector* (**COV**) of the four images considered in the example are shown.

for the typical purposes of a content-based image retrieval (CBIR) system, like image classification or retrieval.

4 Results and Discussion

We used two different datasets to test our approach. The first dataset D1 (412 images) is composed by 5 categories of the Washington database[1], *i.e.*, *Arborgreen*,

[1] http://www.cs.washington.edu/research/imagedatabase/

Green Lake, *Cherry*, *Swiss mountains* and *Greenland*. Such categories present mostly landscape with several FG objects (cars, persons, small buildings). The second database D2 is formed by some categories of the Corel photo CD, as in [3]. We chose $B = 6$ for the OB model inference, and we set $S_{MAX} = 3$.

As first analysis, we observe qualitatively the content of the clusters found by our model on the two datasets. We obtain 8 prototypes for the D1 dataset and 12 for the D2 dataset. The clusters are meaningful, and it is reasonable to assign them explicative concept labels, listed in Fig.4. In order to compare quantitatively the BG concepts extracted by our approach we consider the method for scene classification proposed in [3]. Here, we divide each image of the two datasets in 100 non overlapping patches. We adopt the natural concepts proposed in [3], *i.e.*, *sky*, *water*, *grass*, *trunks*, *foliage*, *field*, *rocks*, *flowers* and *sand* to label each patch. Subsequently, we learn a SVM classifier for each one of the concept, employing a training subset T1 for the dataset D1 and T2 for the dataset D2. We use the classifiers to label the patches of the remaining images. We did similarly using our approach, labeling with our multi-scale method (Sec. 3) all the patches (16 for every image) of testing images with the concepts found by the OB model on T1 and T2. We classify thus the remaining patches. The classification results are shown on Table 1-**A** for what concerns the global accuracy and in Fig.4 for what concerns the confusion matrices. The retrieval experiments have been evaluated by selecting randomly images from the commercial Corel stock photo collection. First of all, we build for each image of our datasets the related COV. For each test image, we evaluate its description given by our model; using such description, we build its COV representation. Then, to retrieve images from our dataset given the test image, we employ the intersection distance as similarity measure.

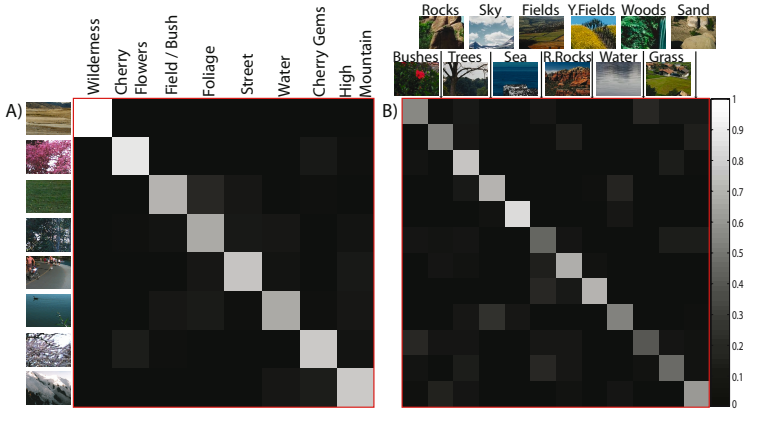

Fig. 4. Concept listings and confusion matrices on A) Washington dataset concepts; B) Corel Dataset concepts. Together with the matrices, in correspondence of each concept the images (or part of them) whose histogram distances from the prototype representing the concept are minimal are shown.

Table 1. A) Image classification results. The classes considered for the Corel dataset are those used in [3]: *Sky/Clouds*, *Forest*, *Fields*, *Lake/Water*, *Coasts*, *Mountains*. **B**) Retrieval performance measures for three different categories of queries.

A) Dataset	OBM	[3]
Washington	84%	79%
Corel	85%	74%

B) Query Type		Rank	$R_P(.5)$	E_R
Easy	*Mountains*	1	0.84	0.26
Medium	*Lake*	1.8	0.77	0.48
Hard	*Coasts*	2.4	0.44	0.63

Like in previous work [5], we draw a graph for the three kind of queries: *easy*, *medium* and *hard queries*. In particular, we choose the category *Mountain* as easy query, *Lake* as medium query, and *Coast* as hard query.

In Table 1-**B**, additional retrieval performance measures are reported in terms of precision/recall graphs. *Rank* is the rank at which the first relevant image is retrieved, $R_P(.5)$ is the recall at precision 0.5 and E_R is the error rate. At the bottom of Fig. 5 some visual examples of the results of the three kind of queries are shown.

In summary, 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 makes it possible to work also with personal pictures typically affected by clutter, i.e., information not useful for scene classification (e.g., persons, faces, etc.). The specific image representation and the multi-scale procedure have been proved capable to obtain good performances for retrieval and classification of images. Further efforts will be planned to analyze more in details the performances of our method on the retrieval task, by testing it on larger standard databases.

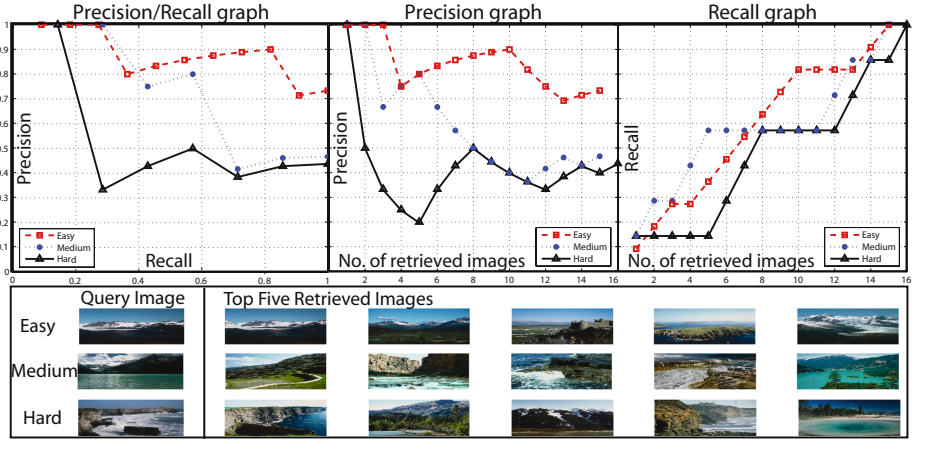

Fig. 5. The precision/recall graph, precision graph, and the recall graph. At the bottom, some visual examples of the results of the three kinds of queries are shown. While the precision represents the fraction of retrieved images that are relevant, the recall represents the fraction of all the relevant images retrieved.

References

1. Henderson, J.: Introduction to real-world scene perception. Visual Cognition 12(3), 849–851 (2005)
2. Torralba, A., Oliva, A.: Statistics of natural image categories. Network: Computation in Neural Systems 14(3), 391–412 (2003)
3. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. Int. J. Comput. Vision 72(2), 133–157 (2007)
4. Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, L.: An introduction to variational methods for graphical models. Machine Learning 37(2), 183–233 (1999)
5. Müller, H., Müller, W., Squire, D.M., Marchand-Maillet, S., Pun, T.: Performance evaluation in content-based image retrieval: Overview and proposals. Pattern Recognition Letters 22(5), 593–601 (2001)