Natural scenes categorization by hierarchical extraction of typicality patterns

A.Perina M.Cristani V.Murino

Dipartimento di Informatica,
Università degli Studi di Verona,
Strada le Grazie 15, 37134 Verona, Italia.
{perina,cristanm,murino}@sci.univr.it

Abstract

Natural scene categorization of images represents a very useful task for automatic image analysis systems in a wide variety of applications. 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 efficiently represent the considered image collection. This vocabulary is formed by a discrete set of visual codewords whose co-occurrence 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 from the scene to be classified or other objects, independent from the scene, eventually present in it. As quoted by perceptual studies, features regarding objects present in an image are not useful to scene categorization, indeed bringing an important source of clutter, in dependence of their size. In this paper, a novel, multiscale, statistical approach for image representation aimed at scene categorization is presented. The method is able to select, at different scales, 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 "natural scene" is usually intended as the one of *a semantically coherent, namable human-scaled view of an outdoor real world environment* [4], and the term "natural scene categorization" refers to the task of grouping images into semantically meaningful categories [18]. Natural scene categorization is without doubts a difficult task and an open research field, because any natural scene category will be characterized by a high degree of diversity and potential ambiguities. Consequently, many efforts are spent on scene categorization, and a huge amount of different approaches are present in literature [18, 10, 7, 13, 8]. In this multifaceted research field, it is widely accepted that methods which provides high categorization accuracies are not the *best methods* in an absolute sense, *i.e.*, high accuracy should not be the primary evaluation criterion for categorization; instead, emphasis should also be given on *how* high accuracy is reached, in order to ensure generalization and robustness.

Typicality-oriented categorization approaches are just focused on these aims, employing categorization measures which implicitly ignore occasional or unexpected visual objects in an image, highlighting viceversa expected (*i.e.*, typical) patterns [18]. This idea finds confirmation by a psychological cognitive perspective, because it is well-known that human relies on global visual properties to exploit scene classification, avoiding to perform recognition of particular objects in the scene [10]. As example, see Fig.1: first two images in the fourth row, extracted randomly from the "Swiss mountains" class from the Washington image database[1], are undoubtedly related because of the presence of *snow-rock* patterns, that typically represent mountains, while the presence of persons plays a marginal role. In this paper, we propose a statistical image representation method for natural scene categorization purposes, which is based on a hierarchical feature extraction scheme. We call this model *hierarchical pattern* (HP) model. Images in a database are first considered as whole entities, trying to organize them taking into account only for frequent and global patterns, disregarding small-size and/or unexpected visual entities. This permits to group together images representing homogeneously typical patterns such as the first two bushes im-

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

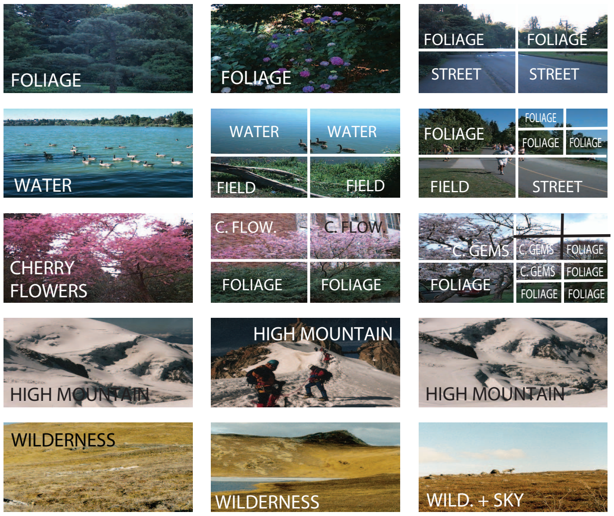

Figure 1. Categorization results on images taken form the test database.

ages depicted in Fig.1, first row. The representation mechanism is further applied at a finer scale to those images which depict partially and/or in combination different *local* typical patterns, obtaining so a structured categorization labeling, as depicted in Fig.1. Our representation scheme is evaluated both qualitatively and quantitatively using cross-validation on standard data, and compared with several state-of-the-art methods.

The rest of the paper is organized as follows. Recent studies of image categorization are briefly reviewed in Section 2. Section 3 gives a general view of the proposed approach, whose details are extensively described in Sections 4 and 5. Finally, Section 6 reports experimental results and concludes the paper.

2 Related work

Methods for scene categorization can be separated in local and global methods. The main hypothesis underlying 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 because features which characterize natural images are very redundant, co-occurrent and therefore robust to clutter.

Local methods increased their importance in these last years, due to the "bag of words" paradigm [5]. Transposed to the image domain, a bag of words becomes a bag of "visterms", i.e. visual features co-occurring in the image [8]. The drawback of these methods is that this representation contains no information about the relative spatial relationship of the visterms. In [11] a local method based on bag of visterms is proposed for scene categorization. Visterms here are cluster centroids of SIFT features [1] found by K-means [3]. Experiments are divided in three separate problems aimed at distinguish indoor/outdoor, city/landscape, sunset/mountain/forest scenes. In [13], bags of visterms have been augmented with a "weak" local spatial modeling. In [7], the idea is to repeatedly subdividing the image and computing histogram of local features at increasingly finer resolution.

Global analysis to perform scene categorization represents the first strategy adopted in the machine learning literature [14, 16, 12]. Recently, Oliva and Torralba propose a general global algorithm for describing images [9], refined more recently in [10]. The key concept is the spatial envelope, which encode five global properties of the scenes (naturalness, openness, roughness, expansion, ruggedness). The approach extracts features from the images power spectrum by convolving it with Gabor-like filters at 12 orientations and 5 scales.

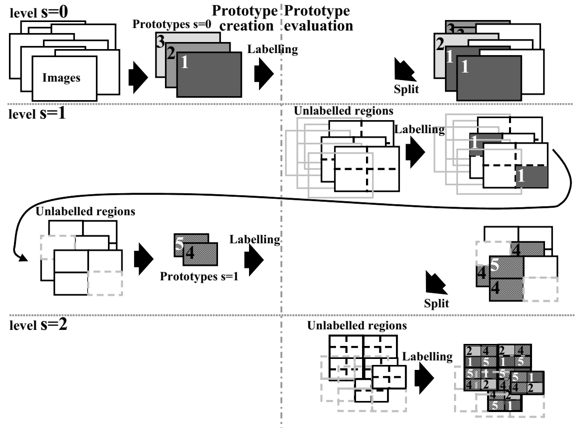

Figure 2. Hierarchical pattern model.

3 Outline of the proposed approach

The proposed HP model analyzes images in a hierarchical fashion with a two-step process, starting from hierarchy detail level $s = 0$, *i.e.*, considering the images as whole entities. The whole process is depicted schematically in Fig.2. The first step is called *prototype creation*; in practice, images are thought as generated by typical prototypes, *i.e.*, patterns that bring only *frequent and global* visual traits of the images, plus random noise. Prototype creation consists in learning via generative modeling such prototypes, and the noise distribution. The second step is called *prototype evaluation*. Here, images are tried to be organized in distinct classes by taking into account for prototypes as class templates. Images class membership is evaluated through an ad-hoc similarity distance. After this, images depicting homogeneously distinct environmental visual patterns (such as the *bushes* patterns exhibited in the first two images of the first row of Fig.2), are grouped together.

If some images remain un-labeled, *i.e.*, they are not glob-

ally valid members of any category (such as the third image in Fig.2, first row), a finer categorization is performed, by dividing them in 4 squared non-overlapping patches (detail level $s = 1$). Therefore, such patches are tried to be labeled using pre-existent global labels, exploiting the intuition that some local patterns could represent global patterns depicted at a smaller scale. Eventual remaining unlabeled patches will be re-organized together, forming novel *local* typical patterns, re-proposing the prototype creation process at a finer scale $s = 1$. In this way, *frequent and local* patterns concur to create local descriptions, which provide structured categorization of an image.
This mechanism of prototype creation and evaluation is performed iteratively by evaluating at level $s = i$ squared patches corresponding to $1/2^i$-*th* of an image, until a user-defined level $s = s_{MAX}$ is reached. As robust image description, we choose a quantized color-histogram in HSV color space. In this way, each image $\mathbf{z}$ is specified by the histogram's bins values $\mathbf{h} = h_1, \dots h_M$, where M varies depending on the quantization selected; in this case we quantize the HSV space with hue $= 20$, saturation $= 4$ and value $= 4$ bins, for a total of $M = 320$ bins. In the following, we intend the image histogram and the image as the same entity, distinguishing them whereas necessary.
Note that the choice of color histogram is functional to show the potentialities of the proposed method, but other, more informative, histogram representations could be used. Purpose of this paper is to show the expressivity of our method using a simple image descriptor.

4 Prototype creation

The goal of this step is to extract from a pool of HSV image histograms a set of B *typical* histogram prototypes, plus a set of B related *typical* masks. Each b-th mask isolates the color bins which are more representative for the b-th prototype. Histogram prototypes are estimated by learning a generative graphical model; each one of them represents a typical image pattern. The generative process that forms an observed image histogram $\mathbf{h}$ is the following one: starting from a prototype class b extracted with probability π_b, then, conditioned on $\mathbf{b}$, a latent binary mask variable $\mathbf{m} = (m_1 \dots m_M)$ parameterized by α_b is chosen. Then, conditioned on the values of the mask and the prototype class, the observed histogram $\mathbf{h}$ is extracted; in practice, whereas the mask bins permit it, that is, whereas $\{m_i\} = 1$, the bins of the observed image histogram are independently extracted with probability following a Gaussian distribution $\mathcal{N}(\mu^{(b)}, \mathbf{\Sigma}^{(b)})$, where $\mathbf{\Sigma}^{(b)} = \sigma^2 \mathbf{I}$. The formal prototype creation generative model is shown in Fig.3 (left). In Fig.3 (right), the generative process is explained using an example taken from the experimental data, and rearranged to fit in the figure (*i.e.* we quantize the HSV color space into M=20 color levels, instead of 320): please note that in the

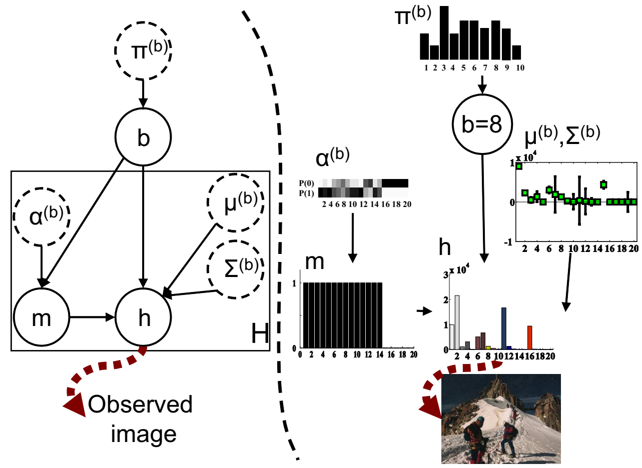

Figure 3. Prototype creation generative model: on the left, the formal model; on the right, the histogram generative process.

image on the bottom, red persons are considered as unexpected patterns by the prototype $b = 8$. In facts, the α_b mask distribution for that class excludes bins that range from the 16-th until the 20-th. Histogram bin values from the 1-th to the 16-th are chosen extracting from the $\mu^{(b)}, \mathbf{\Sigma}^{(b)}$ parameters, shown in the figure error bar (Fig.3). Summarizing, the joint probability of the model is the following:

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

where h_i indicates the *i-th* histogram bin, m_i the *i-th* mask value and M the number of bins of the histogram. Parameterizing the joint we obtain

$$P(\mathbf{h}, \mathbf{m}, \mathbf{b}) = \pi_b \cdot \prod_{i=1}^{M} \alpha_{bi}^{m_i} (1 - \alpha_{bi})^{1-m_i} \mathcal{N}(h_i; \mu_i^{(b)}, \Sigma_i^{(b)})^{m_i} \quad (2)$$

where $\mu_i^{(b)}$ and $\Sigma_i^{(b)}$ are respectively the i-*th* mean and variance values of the b-*th* prototype which model the i-*th* observed histogram bin, and α_{bi} the probability that $m_i = 1$, related to prototype b. The parameters learning has been performed using the Expectation Maximization (EM) algorithm, in a mean field variational version [6]. The learning step converges averagely after 15-20 iterations. In this way, B histogram prototypes can be estimated, by evaluating $\mu_i^{(b)}$ and $\Sigma_i^{(b)}$.

5 Prototype evaluation

The *prototype evaluation* step uses the prototypes found in the former steps, to separate image histograms in disjoint classes. Given an image histogram $\mathbf{h}$, we calculate its

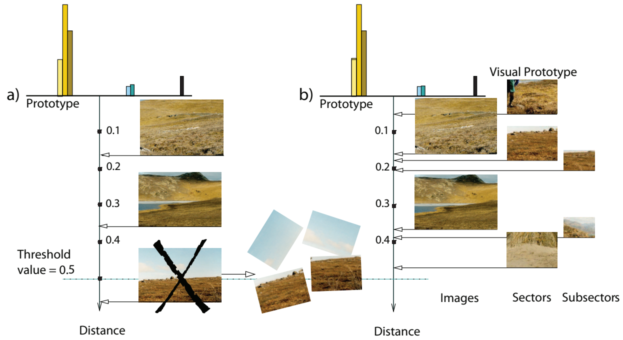

Figure 4. Class (pattern) prototype, visual prototypes and intersection distance.

distance with all the histogram prototypes; the aim is here not only to evaluate the "nearest" prototype, which would give an appropriate class label. Actually, for this task, using simply the likelihood $P(\mathbf{h}|\mathbf{b}) = \sum_{\mathbf{m}} P(\mathbf{h},\mathbf{m}|\mathbf{b}) = P(\mathbf{h}|\mathbf{m},\mathbf{b})P(\mathbf{m}|\mathbf{b})$ would be enough. Instead, we need a distance that 1) gives the nearest prototype; 2) indicates when an histogram $\mathbf{h}$ represents an image which has to be locally further analyzed. Therefore, as similarity distance, we employ the *intersection distance* which is a robust dissimilarity measure among histograms. The intersection distance is evaluated between a test histogram $\mathbf{h}$ and a prototype $\mu^{(b)}$ and it is defined as follows:

$$\mathcal{D}_{Int}(\mu^{(b)}, \mathbf{h}) = 1 - \frac{\sum_i^M \min\left(h_i, {\mu_i}^{(b)}\right)}{|h_i|} \quad (3)$$

The negative term above measures the percentage of the image histogram $\mathbf{h}$ covered by the *b-th* prototype; note that $\mathcal{D}_{Int} \in [0,1]$ [2].

Setting a threshold τ on this distance permit us to 1) choose those images which belong to a particular prototype b, *i.e.*, whose minimal distance $\mathcal{D}$ is realized with the b-th prototype *and* it is below τ; 2) deciding which images need to be split and further analyzed *i.e.*, whose minimal distance $\mathcal{D}$ is realized with the b-th prototype *and* it is above τ. In this paper, we choose $\tau = 0.5$. In this way, we can devise a partition which separates images well modeled by a histogram prototype, forming the set $\{z^{(OK,s=0)}\}$, and images which are not, forming the set $\{z^{(NO,s=0)}\}$, where $s = 0$ indicates the detail level in which images are taken as whole entities. Images in $\{z^{(NO,s=0)}\}$ are thus divided in 4 not-overlapping squared regions, forming the set $\{z^{(NO,s=1)}\}$, where the apex $s = 1$ addresses quarters of images (*i.e.*, histograms calculated on quarters of images). Figure 4a shows the situation after the first prototype evaluation step. All the images whose distance from the prototypes exceeds the threshold τ will be split.

[2] A different distance which considers also the variance of the prototype is currently under study. Anyway, results obtained with the $\mathcal{D}$ distance satisfy us.

Considering each image $\mathbf{z} \in \{z^{(NO,s=1)}\}$, we can evaluate two possibilities: 1) it is modeled by a pre-existent prototype at previous levels (in this case only $s = 0$), or 2) it could represent a novel prototype. To check 1), we up-sample via bi-cubic interpolation the regions $\{z^{(NO,s=1)}\}$ in order to build up-sampled histograms $\{\tilde{h}^{(NO,s=1)}\}$ which are equinumerous with the ones of the previous detail level ($s = 0$).

At this point, the classification via the intersection distance is applied again, still using the threshold value $\tau = 0.5$. Some of the regions can now be assigned to one of the pre-determined global prototypes, enlarging the related class. All the other regions are sent back to the prototype creation process, which can be applied now at a smaller scale $s = 1$; doing this, it will be possible to find a set of region prototypes which can explain local and repetitive aspects of the images.

The process of prototype creation-evaluation is thus reiterated until a smaller scale s_{MAX} is reached, set here to $s_{MAX} = 2$. Note that the last prototype evaluation step does not necessarily assign a class label to all the regions: if a region has a distance value higher than the threshold τ for each prototype it will be labeled as *unknown* sector.

After that, each image could be explained with a structure of typical pattern classes; in order to assign an intuitive meaning to each class, we introduce the *visual prototype* of class b as the region (image, quarter of image, etc.) whose histogram $\mathbf{h}$ is nearest to the prototype $\mu^{(b)}$ (Fig.4b).

6 Experiments and discussions

The images used to train the HP model are taken from the Washington database. Such data-set is divided in 22 categories, originally created for object recognition purposes, so not well-suited for our goals. As in [15, 17], we took only a subset of the database to preserve the consistency with our goals, which are: `Arborgreens`, `Green Lake`, `Cherry`, `Swiss Mountains` and `Greenland` for a total amount of nearly 450 images, similarly as quantity to what done by [17].

At each prototype creation step we learn $B = 5$ prototypes and we set $s_{MAX} = 3$, considering thus $1/16$ of images. Related classes found after the prototype evaluation steps are deleted if they contain less than 5 images. At the end of the HP process, we found 9 prototypes, founded at different detail levels; the related visual prototypes are depicted in Fig.5. As test for the significance of the obtained prototypes, we consider the ground truth semantic annotations of the Washington database, 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 cases examined (see Tab.6), *i.e.*, in 100 times we find an image in which a region belonging to class 1 is present, in the annotations of that image we found at least

Table 1. Correspondence between concepts.

HP Model prototypes	Washington annotations
Class 1	Frozen lake, Grass
	Mountain, Rocks, Ground.
Class 2	Cherry tree, Tree, Trunk
Class 3	Frozen lake, Grass
	Clear Sky, Cloudy sky, Clouds
	Partially cloudy sky
Class 4	Trees, Bushes, Trunk, Fern, Lilly
Class 5	Grass, Bush, Flowers, Ground.
Class 6	Snow, Rockes, Ice, Clear Sky
Class 7	Water
Class 8	Street, Trail, Sidewalk
Class 9	Cherry tree, Flowers

90 times the "Frozen lake" and "Grass" annotations. The typicality of the classes found is evident: therefore, we label the classes as, respectively: *Wilderness, Cherry Flowers, Wilderness + Sky, Foliage, Field/Bush, High Mountain, Water, Street* and *Cherry Gems*. We then apply these labels to each image of the database, creating a sort of rough segmentation, in which *natural* scene labels are given, without taking into account for unexpected patterns (see Fig.1). In

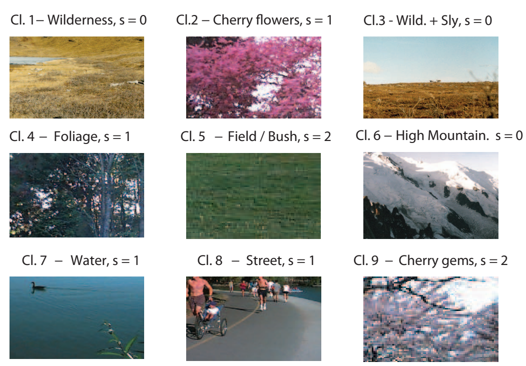

Figure 5. Visual prototypes.

order to evaluate the *peculiarity* of the typical patterns, we create a "ground truth" database, by manually labeling all the images of our database, using the typical pattern labels over regions large 1/16 of image. Then, we repeatedly extract a small training set (10% of the total images) using the un-labeled images, with which we learn the HP model; subsequently, we evaluate the correspondences of the labels given by our approach with those given manually. The resulting classification table is depicted in Tab.2.

6.1 Comparison with a semantic modeling with SIFT and pLSA

In [2] outperforming *scene classification results* are presented, obtained using pLSA and color SIFT features.

Table 2. Confusion matrix representing the concepts classification.

Overall 77%	Classification in %							
	wi	*cf*	*fi*	*fo*	*st*	*wa*	*cg*	*hm*
True Class								
Wilderness	**100**	0	0	0	0	0	0	0
Cherry flowers	0	**88**	0	0	0	0	10	2
Field / Bush	0	0	**69**	19	9	0	3	0
Foliage	1	2	6	**66**	10	9	0	6
Streets	0	0	0	9	**76**	5	0	10
Water	0	0	8	12	2	**67**	2	9
Cherry gems	0	14	2	0	0	0	**78**	6
High Mountain	0	0	0	0	0	9	14	**77**
Unknown	5	0	5	3	13	28	6	43

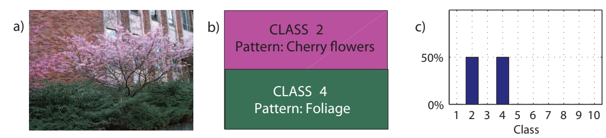

Figure 6.

Briefly speaking, with pLSA each image $\mathbf{z}$ is modeled as a mixture of *topics* (MOT) $\mathbf{t} = t_1, \ldots, t_B$. We do the same with our typical patterns, creating the *pattern-occurrence vector* (**POV**) which is essentially a normalized histogram of the class (pattern) occurrences in an image (Fig.6c). In order to compare our approach with [2], we run pLSA using $B = 10$ topics, using the same features used in that paper. Then, for each image, we have a particular MOT and POV. In order to discover how much is informative such characterization, we calculate the *category* **POV**, *i.e.*, the average POV built by considering all the POV of a given category, which is

$$p^c = \frac{1}{N_c} \sum_{j=1}^{N_c} \mathbf{POV}(j) \qquad (4)$$

where c refers to one of the five Washington categories considered, and N_c is the number of images in that category. The do the same with the MOT's, building the *category* **MOT**. Figure 7 displays the *category* **POV** and **MOT** with the related standard deviations for each typical pattern/topic c. The figure easily reveals which semantic concepts are especially discriminant for each category. In particular, regarding the POVs, every category has it own characteristic pattern: *foliage* for Arborgreens, *field/bush* and *water* for Green lake, *cherry flowers* for Cherries, *high mountain* for Swiss mountains and *wilderness* for Greenland. Such peaky trend is absent in the MOTs. The reason stays at "feature level". While SIFT features are the optimum for object or

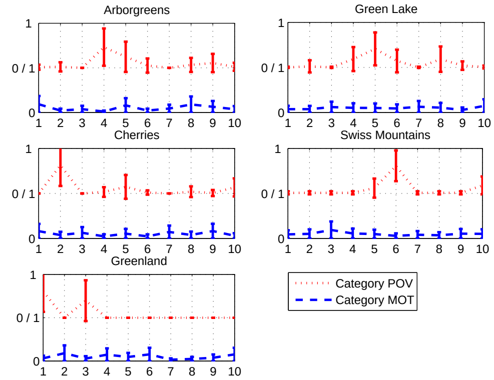

Figure 7. Prototypes and standard deviations of the five scene categories.

general scene recognition, for natural scene recognition or image semantical analysis they provide poor performance as they try to classify an image through the "visual objects" that appear within (Fig.8), such as cars, persons or animals, unimportant for natural scene categorization. As quantitative test of the category description ability of our method, we calculate the confusion matrices of the *category* **POV** and **MOT**, calculating the inter category intersection distance, normalized to $[0, 1]$, among *category* **POV** (Fig.9, on the left) and **MOT** (Fig.9, on the rigth).

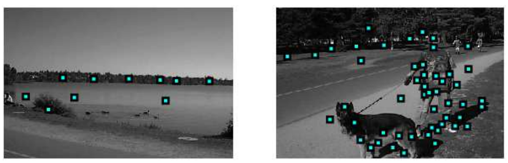

Figure 8. Two images and the SIFT features identified marked with a square.

	AG	GL	CH	SM	GR
AG					
GL	0,38				
CH	0,76	0,77			
SM	0,78	0,71	0,74		
GR	0,96	1	1	1	

	AG	GL	CH	SM	GR
AG					
GL	0,24				
CH	0,32	0,29			
SM	0,39	0,37	0,34		
GR	0,31	0,28	0,35	0,39	

Figure 9.

References

[1] *Object recognition from local scale-invariant features*, volume 2, 1999.

[2] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *Proceedings of the ECCV*, 2006.

[3] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, second edition, 2001.

[4] J. Henderson. Introduction to real-world scene perception. *Visual Cognition*, 12:849–851(3), August 2005.

[5] 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*, pages 50–57, New York, NY, USA, 1999. ACM Press.

[6] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

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

[8] F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 2005 CVPR*.

[9] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, 2001.

[10] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research: Visual perception*, 155:23–36, 2006.

[11] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 883–890, Washington, DC, USA, 2005. IEEE Computer Society.

[12] Y. Rui, T. Huang, and S. Chang. Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10(4):39–62, 1999.

[13] J. Sivic, B. Russell, A. Efros, A. Zisserman, and B. Freeman. Discovering objects and their location in images. In *International Conference on Computer Vision (ICCV 2005)*, October 2005.

[14] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[15] O. van Kaick and G. Mori. Automatic classification of outdoor images by region matching. In *CRV '06: Proceedings of the The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, page 9, Washington, DC, USA, 2006. IEEE Computer Society.

[16] R. Veltkamp, M. Tanase, and D. Sent. Features in content-based image retrieval systems: a survey. In *State-of-the-Art in Content-Based Image and Video Retrieval*, pages 97–124. Kluwer, 1999.

[17] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *International Journal of Computer Vision*. (in press).

[18] J. Vogel and B. Schiele. A semantic typicality measure for natural scene categorization. In *DAGM-Symposium*, pages 195–203, 2004.