

# Scene Classification Using a Hybrid Generative/Discriminative Approach

Anna Bosch, Andrew Zisserman, and Xavier Muñoz

**Abstract**—We investigate whether dimensionality reduction using a latent generative model is beneficial for the task of weakly supervised scene classification. In detail, we are given a set of labeled images of scenes (for example, coast, forest, city, river, etc.), and our objective is to classify a new image into one of these categories. Our approach consists of first discovering latent “topics” using probabilistic Latent Semantic Analysis (pLSA), a generative model from the statistical text literature here applied to a bag of visual words representation for each image, and subsequently, training a multiway classifier on the topic distribution vector for each image. We compare this approach to that of representing each image by a bag of visual words vector directly and training a multiway classifier on these vectors. To this end, we introduce a novel vocabulary using dense color SIFT descriptors and then investigate the classification performance under changes in the size of the visual vocabulary, the number of latent topics learned, and the type of discriminative classifier used (k-nearest neighbor or SVM). We achieve superior classification performance to recent publications that have used a bag of visual word representation, in all cases, using the authors’ own data sets and testing protocols. We also investigate the gain in adding spatial information. We show applications to image retrieval with relevance feedback and to scene classification in videos.

**Index Terms**—Scene classification, pLSA, spatial information.

## 1 INTRODUCTION

CLASSIFYING scenes (such as mountains, forests, and offices) is not an easy task owing to their variability, ambiguity, and the wide range of illumination and scale conditions that may apply. As was noted in [3], two basic strategies can be found in the literature. The first uses low-level features such as global color or texture histograms, the power spectrum, etc., and is normally used to classify only a small number of scene categories (indoor versus outdoor, city versus landscape, etc.) [29], [30]. The second strategy uses an intermediate representation before classifying scenes [9], [24], [32] and has been applied to cases where there are a larger number of scene categories (up to 15).

In this paper, we follow the second strategy and introduce a classification algorithm based on a combination of unsupervised probabilistic Latent Semantic Analysis (pLSA) [16] followed by a discriminative classifier. The pLSA model was originally developed for topic discovery in a text corpus, where each document is represented by its word frequency. Here, it is applied to images represented by the frequency of “visual words” [28]. The formation and performance of this “visual vocabulary” is investigated in depth. In particular, we compare sparse and dense feature descriptors over a number of modalities (color, texture, and orientation). The approach is inspired in particular by three previous papers: 1) the use of pLSA on sparse features for recognizing compact object categories (such as Caltech cars and faces) in Sivic et al. [27],

2) the dense SIFT-like [21] features developed in Dalal and Triggs [8] for pedestrian detection, and 3) the semisupervised application of Latent Dirichlet Analysis (LDA) for scene classification in Fei-Fei and Perona [9]. We have made extensions over all three of these papers both in developing new features and in the classification algorithm. Our work is most closely related to that of Quelhas et al. [25] who also use a combination of pLSA and supervised classification. However, their approach differs in using sparse features and is applied to classify images into only three scene types.

We compare our classification performance to that of four previous methods [9], [18], [24], [32] using the authors’ own databases. This previous work uses varying levels of supervision in training (compared to the unsupervised topic discovery developed in this paper): Fei-Fei and Perona [9] requires the category of each scene to be specified during learning (in order to discover the *themes* (topics) of each category)—we do not specify the category when discovering topics; Oliva and Torralba [24] requires a manual ranking of the training images into six different properties; and Vogel and Schiele [32] requires a manual classification of 60 K local patches from the training images into one of nine *semantic concepts*. As will be seen, we achieve superior performance in all cases. Lazebnik et al. [18] do not use an intermediate topic representation but improve performance by adding spatial information over the bag of words model. We compare a number of methods that include both latent models and spatial information and demonstrate improved results over [18].

We briefly give an overview of the pLSA model in Section 2. Then, in Section 3, we describe the hybrid classification algorithm based on applying pLSA to images followed by discriminative classification. Section 4 describes the features used to form the visual vocabulary and the principal parameters that are investigated. A description of data sets and a detailed description of the experimental

- A. Bosch and X. Muñoz are with the Computer Vision and Robotics Group, Universitat de Girona, Campus Montilivi, Edifici P-IV, Av. Lluís Santaló, s/n, 17071, Girona, Spain. E-mail: {aboschr, xmunoz}@eia.udg.es.
- A. Zisserman is with the Robotics Research Group, University of Oxford, Parks Road, Oxford OX1 3PJ, UK. E-mail: az@robots.ox.ac.uk.

Manuscript received 15 Sept. 2006; revised 18 Feb. 2007; accepted 8 May 2007; published online 8 June 2008.

Recommended for acceptance by M. Hebert.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0663-0906. Digital Object Identifier no. 10.1109/TPAMI.2007.70716.

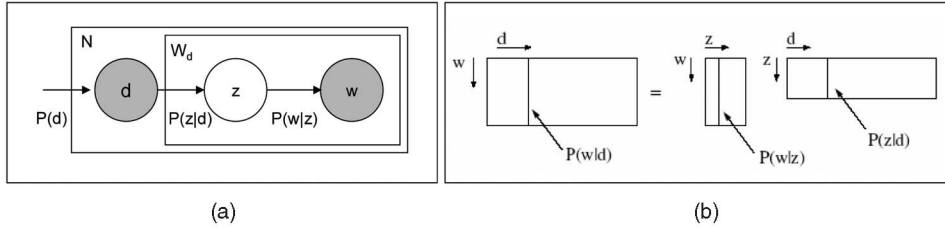


Fig. 1. (a) pLSA graphical model. Nodes inside a given box (plate notation) indicate that they are replicated the number of times indicated in the top left corner ( $N$  = number of images;  $W_d$  = number of (visual) words per image). Filled circles indicate observed random variables; unfilled are unobserved. (b) The goal is to find the topic specific word distributions  $P(w|z)$  and corresponding document specific mixing proportions  $P(z|d)$  which make up the observed document specific word distribution  $P(w|d)$ .

procedure is given in Section 5. Section 6 reports the principal investigation of the paper—first, we optimize the performance over changes in the vocabulary and number of latent topics, then, we compare the hybrid classifier to a more standard approach of classifying on the bag of words histograms directly. Section 7 then introduces three models that include spatial information and compares their performance to the model of Lazebnik et al. [18]. In Section 9, we demonstrate applications of the hybrid algorithm to relevance feedback, scene classification in videos, and segmentation. In Section 10, we discuss the ambiguities and difficulties of the scene classification task.

This paper is an expanded version of that in [4]. The extensions include the comparison of the classifiers (K-Nearest Neighbor and SVM), Section 7 on spatial information, and evaluations on new data sets (that in [18] and Caltech 101 [20]).

## 2 PLSA MODEL

Probabilistic Latent Semantic Analysis (pLSA) is a generative model from the statistical text literature [16]. In text analysis, this is used to discover topics in a document using the bag of words document representation. Here, we have *images* as *documents*, and we discover *object categories* as *topics* (for example, grass and houses) so that an image containing instances of several objects is modeled as a mixture of topics. The models are applied to images by using a *visual* analog of a *word*, formed by vector quantizing color, texture, and SIFT feature like region descriptors (as described in Section 4). pLSA is appropriate here because it provides a correct statistical model for clustering in the case of multiple object categories per image. We will explain the model in terms of images, visual words, and topics.

Suppose we have a collection of images  $D = d_1, \dots, d_N$  with words from a visual vocabulary  $W = w_1, \dots, w_V$ . The data is a  $V \times N$  co-occurrence table of counts  $N_{ij} = n(w_i, d_j)$ , where  $n(w_i, d_j)$  denotes how often the term  $w_i$  occurred in an image  $d_j$ . A latent variable model associates an unobserved topic variable  $z \in Z = z_1, \dots, z_Z$  with each observation, an observation being the occurrence of a word in a particular image  $(w_i, d_j)$ . We introduce the following probabilities:  $P(d_j)$  denotes the probability of observing a particular image  $d_j$ ,  $P(w_i|z_k)$  denotes the conditional probability of a specific word conditioned on the unobserved topic variable  $z_k$ , and, finally,  $P(z_k|d_j)$  denotes an image specific probability distribution over the latent variable space. Using these definitions, the generative model is the following:

- Select an image  $d_j$  with probability  $P(d_j)$ .
- Pick a latent topic  $z_k$  with probability  $P(z_k|d_j)$ .
- Generate a word  $w_i$  with probability  $P(w_i|z_k)$ .

As a result, one obtains an observation pair  $(w_i, d_j)$ , where the latent topic variable  $z_k$  is discarded.

The graphical model representation is shown in Fig. 1a corresponding to a joint probability  $P(w, d, z) = P(w|z)P(z|d)P(d)$ . Marginalizing out the latent variable  $z$  gives

$$P(w, d) = \sum_{z \in Z} P(w, d, z) = P(d) \sum_{z \in Z} P(w|z)P(z|d)$$

and then from  $P(w, d) = P(d)P(w|d)$ , we obtain  $P(w|d)$  as

$$P(w|d) = \sum_{z \in Z} P(w|z)P(z|d). \quad (1)$$

This amounts to a matrix decomposition, as shown in Fig. 1b with the constraint that both the topic vectors  $P(w|z)$  and mixture coefficients  $P(z|d)$  are normalized to make them probability distributions. Essentially, each image is modeled as a mixture of topics, the histogram for a particular document being composed from a mixture of the histograms corresponding to each topic. In particular, each image is a convex combination of the  $Z$  topic vectors.

Following the likelihood principle, one determines  $P(w|z)$  and  $P(z|d)$  by maximization of the log likelihood function:

$$L = \log P(D, W) = \sum_{d \in D} \sum_{w \in W} n(w, d) \log P(w, d). \quad (2)$$

This is equivalent to minimizing the Kullback-Leibler divergence between the measured empirical distribution and the fitted model. The model is fitted using the Expectation Maximization (EM) algorithm, as described in [16]. Fitting the model involves determining the topic vectors that are common to all documents and the mixture coefficients that are specific for each document. The goal is to determine the model that gives high probability to the distribution of visual words that appear in the corpus.

## 3 HYBRID CLASSIFICATION

Training proceeds in two stages: First, the topic specific distributions  $P(w|z)$  are learned from the set of training images. Determining both  $P(w|z)$  and  $P(z|d_{train})$  simply involves fitting the pLSA model to the entire set of training images. In particular, it is not necessary to supply the identity of the images (that is, which category they are in) or any region segmentation. Each training image is then represented by a  $Z$ -vector  $P(z|d_{train})$ , where  $Z$  is the number of topics learned.

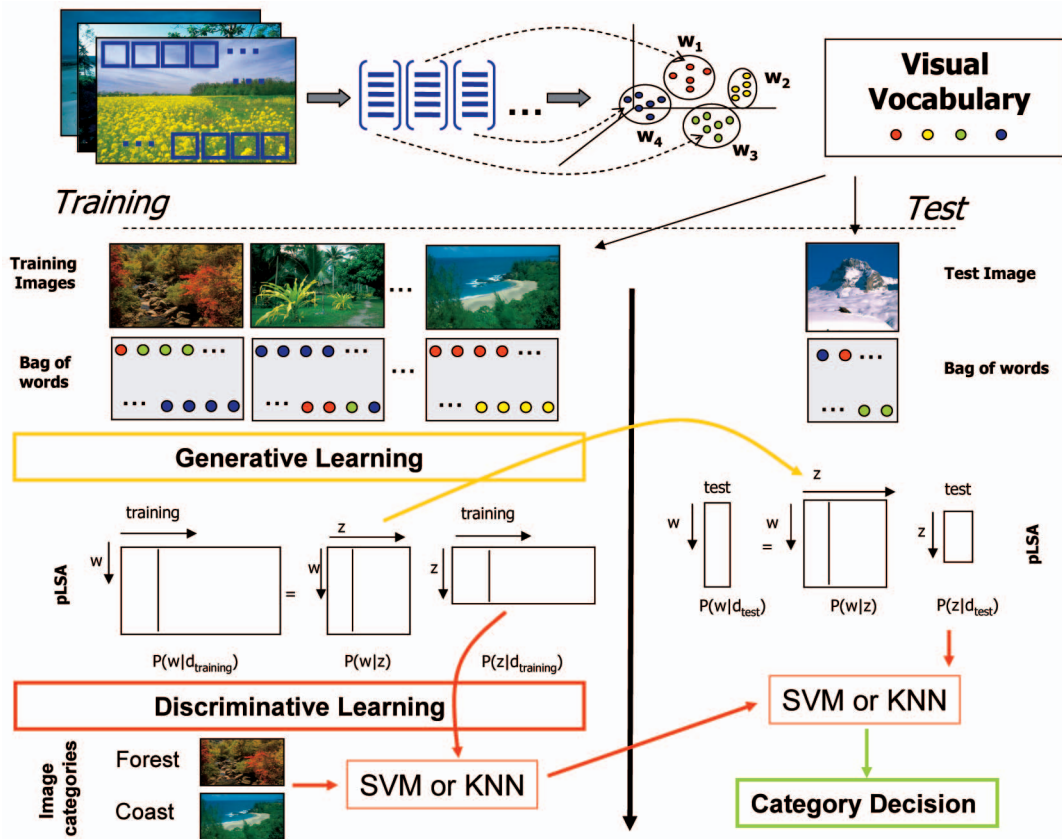


Fig. 2. Overview of visual vocabulary formation, learning, and classification stages.

In the second stage, a multiclass discriminative classifier is trained given the vector  $P(z|d_{train})$  of each training image and its class label. For the discriminative stage, we compare the KNN classifier to an SVM classifier. In more detail, the KNN selects the  $K$  nearest neighbors of the new image within the training database (using euclidean distance). Then, it classifies the test image according to the category label, which is most represented within the  $K$  nearest neighbors. For the SVM classifier, an exponential kernel of the form  $\exp -\alpha d$  is used, where  $d$  is the euclidean distance between the vectors, and the scalar  $\alpha$  is determined as described in [36] (we use the LIBSVM package [5] with the trade-off between training error and margin at  $C = 1$ ). The multiway classification is done using the one-versus-all rule: A classifier is learned to separate each class from the rest and a test image is assigned the label of the classifier with the highest response.

Classification of an unseen test image similarly proceeds in two stages: First, the document specific mixing coefficients  $P(z|d_{test})$  are computed, and these are then used to classify the test images using a discriminative classifier. In more detail, document specific mixing coefficients  $P(z|d_{test})$  are computed using the fold-in heuristic described in [15]. The unseen image is projected onto the simplex spanned by the  $P(w|z)$  learned during training, that is, the mixing coefficients  $P(z_k|d_{test})$  are sought such that the Kullback-Leibler divergence between the measured distribution and  $P(w|d_{test}) = \sum_{z \in Z} P(w|z)P(z|d_{test})$  is minimized. This is achieved by running EM in a similar manner to that used in learning, but now, only the coefficients  $P(z_k|d_{test})$  are updated in each M-step with the learned  $P(w|z)$  kept fixed. The result is that the test image is represented by a  $Z$ -vector. The test image is

then classified by the multiclass discriminative classifier (KNN or SVM), as described above. Fig. 2 shows graphically the hybrid generative/discriminative process for both training and testing.

#### 4 VISUAL WORDS AND VISUAL VOCABULARY

In the formulation of pLSA, we compute a co-occurrence table, where each image is represented as a collection of visual words, provided from a visual vocabulary. This visual vocabulary is obtained by vector quantizing descriptors computed from the training images using k-means, see the illustration at the first part in Fig. 2. Previously, both sparse [7], [17], [28] and dense descriptors, for example, [8], [19], [31] have been used. Here, we carry out a thorough comparison over dense descriptors for a number of visual measures (see below) and compare to a sparse descriptor. We vary the size of the patches and degree of overlap and compare normalized to unnormalized images. We then assess classification performance over four different image data sets described in Section 5.

We investigate four dense descriptors and compare their performance to a previously used sparse descriptor. In the dense case, the important parameters are the size of the patches ( $N$ ) and their spacing ( $M$ ), which controls the degree of overlap.

**Gray patches (dense).** As in [31] and using only the gray-level information, the descriptor is an  $N \times N$  square neighborhood around a pixel. The pixels are row reordered to form a vector in an  $N^2$  dimensional feature space. The patch size tested are  $N = 5, 7, \text{ and } 11$ . The patches are spaced by  $M$  pixels on a regular grid. The patches do not

overlap when  $M = N$  and do overlap when  $M = 3$  (for  $N = 5, 7$ ) and  $M = 7$  (for  $N = 11$ ).

**Color patches** (dense). Same as above but the color information is used for each pixel. We consider the three color components HSV and obtain an  $N^2 \times 3$ -dimensional vector. As in the gray level, we used  $N = 5, 7$ , and  $11$ . We use HSV because of its similarities to the way humans tend to perceive color and because it is less sensitive to shadow and shading.

**Gray SIFT** (dense). SIFT descriptors [21] are computed at points on a regular grid with spacing  $M$  pixels, here,  $M = 5, 10$ , and  $15$ . At each grid point, SIFT descriptors are computed over circular support patches with radii  $r = 4, 8, 12$ , and  $16$  pixels. Consequently, each point is represented by  $n$  SIFT descriptors (where  $n$  is the number of circular supports), each is 128 dimensional. Multiple descriptors are computed to allow for scale variation between images. The patches with radii  $8, 12$ , and  $16$  overlap. Note, the descriptors are rotation invariant.

**Color SIFT** (dense). Same as above but, now, SIFT descriptors are computed for each HSV component. This gives a  $128 \times 3$  dimensional SIFT descriptor for each point. Note, this is a novel feature descriptor. It captures the color gradients (or edges) of the image. Other ways of using color with SIFT features have been proposed in [12], [34].

**Gray SIFT** (sparse). Affine covariant regions are computed for each gray-scale image, constructed by elliptical shape adaptation about an interest point [22]. These regions are represented by ellipses. Each ellipse is mapped to a circle by appropriate scaling along its principal axis and a 128-dimensional SIFT descriptor computed. This is the method used in [7], [17], [27], [28].

## 4.1 Implementation Details

### 4.1.1 Dense SIFT Descriptors

In most previous applications, SIFT-like descriptors are used following a sparse feature detection and so have only been applied at image points where there is sufficient structure (for example, a strong response from a Harris or Hessian operator). In our case, the SIFT descriptors are applied densely, perhaps at every pixel, and this raises two areas of concern.

First, in regions with near constant color/brightness (like sky, road) that consequently have small image gradients, is the resulting description (the visual words) very sensitive to noise? In practice, we find that the assigned word for such patches is often the same and relatively insensitive to patch size. For example, if sky patches with  $r = 4$  are assigned the word  $w_1$ , then sky patches with  $r = 8$  are also assigned the word  $w_1$ , and so on. Where the small gradients (noise) do result in different random visual word assignments, then the pLSA topic learns this distribution.

Second, is there a problem with noise causing wrap-around in the H color channel? This could occur with a region consisting of small fluctuations around saturated red and would result in an alternation of visual word assignment over that region. However, in practice, we do not observe this problem in the current databases.

### 4.1.2 Normalization

Gray-level images are normalized to have intensities with mean zero and unit standard deviation. Color images are first normalized as in "Gray World" [6], [11] to have R, G, and B components  $R * (\mu/\mu_r)$ ,  $G * (\mu/\mu_g)$ ,  $B * (\mu/\mu_b)$ , where

$\mu = (\mu_r + \mu_g + \mu_b)/3$ , and  $\mu_r, \mu_g$ , and  $\mu_b$  are the mean of each component. The HSV is then computed from these normalized values.

## 5 DATA SETS AND METHODOLOGY

### 5.1 Data Sets

We evaluated our classification algorithm on four different data sets:

1. Oliva and Torralba [24],
2. Vogel and Schiele [32],
3. Fei-Fei and Perona [9], and
4. Lazebnik et al. [18].

We will refer to these data sets as OT, VS, FP, and LSP, respectively. Fig. 3 shows example images from each data set, and the contents are summarized here.

**OT.** Includes 2,688 images classified as eight categories: 360 coasts, 328 forest, 374 mountain, 410 open country, 260 highway, 308 inside of cities, 356 tall buildings, and 292 streets. The average size of each image is  $250 \times 250$  pixels.

**VS.** Includes 702 natural scenes consisting of six categories: 144 coasts, 103 forests, 179 mountains, 131 open country, 111 river, and 34 sky/clouds. The size of the images is  $720 \times 480$  (landscape format) or  $480 \times 720$  (portrait format). Every scene category is characterized by a high degree of diversity and potential ambiguities since it depends strongly on the subjective perception of the viewer.

**FP.** Contains 13 categories and is only available in gray scale. This data set consists of the 2,688 images (eight categories) of the OT data set plus: 241 suburb residence, 174 bedroom, 151 kitchen, 289 living room, and 216 office. The average size of each image is approximately  $250 \times 300$  pixels.

**LSP.** Contains 15 categories and, as with FP, is only available in gray scale. This data set consists of the 13 categories of the FP data set plus: 315 store and 311 industrial. The average size of each image is approximately  $250 \times 300$  pixels.

### 5.2 Methodology

The classification task is to assign each test image to one of a number of categories. The performance is measured using a confusion table, and the overall performance rates are measured by the average value of the diagonal entries of the confusion table.

Data sets are split randomly into two separate sets of images, half for training and half for testing. From the training set, we randomly select 100 images to form a validation set. This validation set is used to find the optimal parameters, and the rest of the training images are used to compute the vocabulary and pLSA topics. A vocabulary of visual words is learned from about 30 random training images of each category.

Excluding the preprocessing time of feature detection and visual vocabulary formation, it takes about 20 minutes to fit the pLSA model to 1,600 images (Matlab implementation on a 1.7 GHz computer).

The new classification scheme is compared to two baseline methods. These are included in order to gauge the difficulty of the various classification tasks. The baseline algorithms are as follows.

**Global color model.** The algorithm computes the global HSV histograms for each training image. The color values are represented by a histogram with 36 bins for  $H$ , 32 bins for  $S$ ,

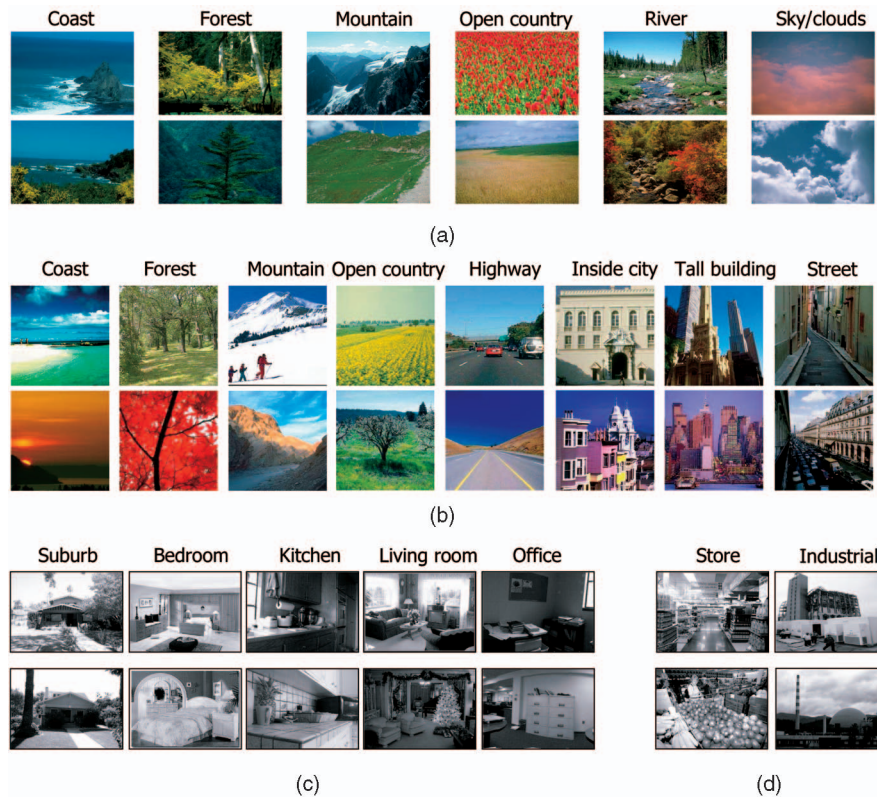


Fig. 3. Example images from the four different data sets used. (a) From data set OT [24], (b) from data set VS [32], (c) from the data set FP [9], the remaining images of this data set are the same as in OT but in gray scale, and (d) from data set LSP [18], same scenes as in FP plus store and industrial.

and 16 bins for  $V$ , giving an 8-dimensional vector for each image. A test image is classified using KNN (with  $K = 10$ ).

**Global texture model.** The algorithm computes the orientation of the gradient at each pixel for each HSV channel at each training image. These orientations are collected into a 72 bin histogram for each color channel and concatenated to form a histogram of  $72 \times 3$  bins for each image. The classification of a test image is again carried out using KNN.

## 6 CLASSIFICATION RESULTS

In this section, we carry out a set of experiments to investigate the various choices of vocabularies, parameters, and classifiers and also to assess the benefits or otherwise of using pLSA as an intermediate representation.

The experiments in this section are all on the OT data set. The results for the other data sets (FP, VS, and LSP) are given in Section 8. For the OT data set, three classification situations are considered: Classification into eight categories, and classification within the two subsets of natural (four categories) and man-made (four categories) images. The latter two are the situations considered in [24].

We start by finding the optimal parameters ( $V$ ,  $Z$ , and  $K$ ) over the validation set for each of the different vocabularies described in Section 6.1. The optimal parameters are then fixed and subsequent results reported on the test set in Section 6.2.

### 6.1 Optimizing the Parameters $V$ , $Z$ , and $K$ (on the Validation Set)

We first investigate how classification performance (on the validation set) is affected by the various parameters: the

number of visual words ( $V$  in the K-Means vector quantization), the number of topics ( $Z$  in pLSA), and the number of neighbors ( $K$  in KNN). Fig. 4 shows this performance variation for two types of descriptors—dense color SIFT with  $M = 10$  and four circular supports, and gray patches with  $N = 5$  and  $M = 3$ . Note the mode in the graphs of  $V$ ,  $Z$ , and  $K$  in both cases. This is quite typical across all types of visual words, though the position of the modes vary slightly. For example, using color SIFT, the mode is at  $V = 1,500$  and  $Z = 25$ , whereas for gray patches, the mode is at  $V = 700$  and  $Z = 23$ . For  $K$ , the performance increases progressively until  $K$  is between 7 and 12 and then drops off slightly.

For color patches, the best performance is obtained when using the  $5 \times 5$  patch over normalized images, with  $M = 3$ ,  $V = 900$ ,  $Z = 23$ , and  $K = 10$ . The best results overall are obtained with dense color SIFT with four circular supports,  $M = 10$ , normalized images,  $V = 1,500$ ,  $Z = 25$ , and  $K = 10$ . We will see in next section that this vocabulary is also the one that gives the best results on the test set.

To investigate the statistical variation, we repeat the dense color SIFT experiment ( $r = 4, 8, 12, 16$  and  $M = 10$ ) 15 times with varying random selection of the training, validation, and test sets, and building the visual vocabulary afresh each time. All parameters are fixed with the number of visual words  $V = 1,500$ , the number of topics  $Z = 25$ , and the number of neighbors  $K = 10$ . We obtained performance values between 79 percent and 86 percent with a mean of 84.7 percent and a standard deviation of 1.9 percent on the test set.

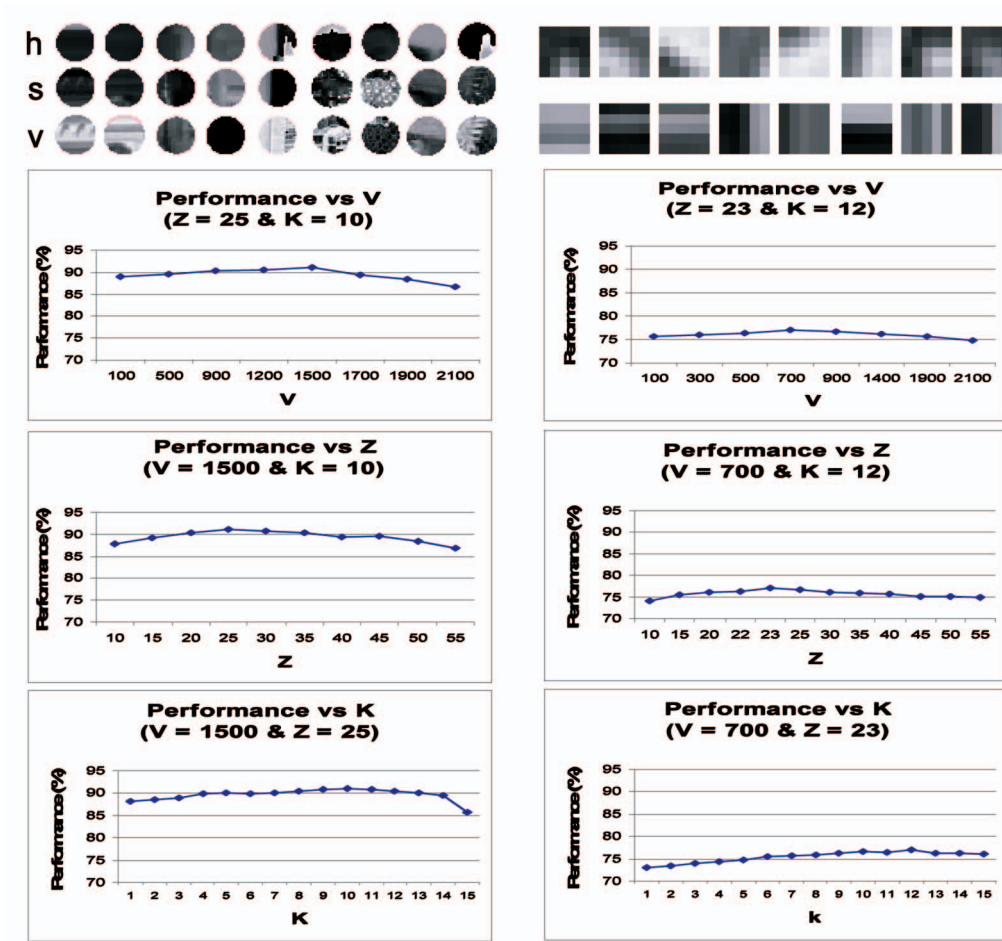


Fig. 4. Validation set performance under variation in various parameters for the eight category OT classification. Left: Example visual words and performance for dense color SIFT  $M = 10, r = 4, 8, 12, 16$  (each column shows the HSV components of the same word). Right: Example visual words and performance for gray patches with  $N = 5$  and  $M = 3$ . Top graph: varying number of visual words,  $V$ . Middle graph: varying number of topics,  $Z$ . Bottom graph: varying  $k$  (KNN).

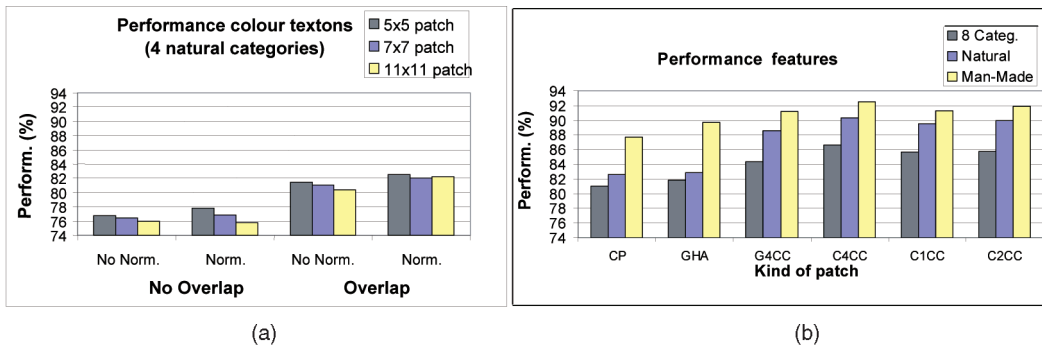


Fig. 5. (a) The OT test set performance when classifying the four natural categories using normalized and unnormalized images and with overlapping and nonoverlapping patches. Color patches are used. (b) Performance when classifying all categories, man-made and natural using different patches and features. Abbreviations for this and subsequent figures: Color Patches, Gray Harris Affine (GHA)—sparse, all the other descriptors are dense, Gray SIFT four Concentric Circles (G4CC), Color SIFT with  $n$  Concentric Circles ( $C_nCC$ ).

## 6.2 Comparison of Features and Support Regions (on the Test Set)

We next investigate the patch descriptors in more detail. Again, we use the OT data set with eight categories and the KNN classifier for this task (the SVM classifier is investigated in Section 6.3). In the following results, the optimum choice of parameters determined on the validation set is used for each descriptor type but here applied to the test set. Fig. 5a shows the results when classifying the images of natural scenes with

color patches. The performance when using normalized images is nearly 1 percent better than when using unnormalized ones. When using overlapping patches, the performance increases by almost 6 percent compared to no overlap. Similar results occur for the man-made and all scene category sets. Comparing results when classifying the images using only gray-level information or using color, it can be seen in Fig. 5b and Table 2 that color brings an increment of around 2 percent. This is probably because color is such an important

TABLE 1

Test Set Performance when Changing the Number of Training and Test Support Regions (for OT Eight Categories)

Training Regions	Testing Regions		
	4CC	2CC	1CC
4CC	<b>86.9</b>	86.7	86.3
Same as Testing	<b>86.9</b>	85.8	85.7

First row: each pixel in the training images are represented by four circles (4CC) and the testing images are represented by four (4CC), two (2CC), and one (1CC) circle from left to right. Second row: pixels in the training and testing images are represented by the same number of circles. The color SIFT descriptor is used.

TABLE 2

Test Set Performance for Different Features when Using the OT Database

Visual Vocabulary	GP	CP	G4CC	C4CC	PS	GIC	GIT
All categ.	71.5	77.0	84.3	<b>86.6</b>	82.6	55.1	64.6
Natural categ.	75.4	82.4	88.5	<b>90.2</b>	84.0	59.5	70.1
Man-made categ.	77.4	83.5	91.1	<b>92.5</b>	89.3	66.1	73.8

Gray Patches (GP), Color Patches (CP), Gray SIFT four Concentric Circles (G4CC), Color SIFT with four Concentric Circles (C4CC), Color Patches and Color SIFT (PS), Global Color (GIC), and Global Texture (GIT).

factor in outdoor images and helps to disambiguate and classify the different objects in the scene.

The performance of SIFT features is shown in Fig. 5b. The best results are obtained with dense and not sparse descriptors. This is almost certainly because we have more information on the images: In the sparse case, the only information is where a Harris detector fires and, especially, for natural images, this is a very impoverished representation. Again, color is a benefit with better results obtained using color than gray SIFT. The performance using gray SIFT when classifying natural images is 88.5 percent and increase 2 percent when using color SIFT both with four concentric support regions. The difference when using these vocabularies with man-made images is not as significant. This reiterates that color in natural images is very important for classification.

**Number of support regions.** Turning to the performance variation with the number of support regions for dense SIFT. It can be seen in Fig. 5b that best results are obtained using four concentric circles. With only one support region to represent each patch, results are around 1 percent worse. This is probably because of the lack of invariance to scale changes: Using four support regions to represent each pixel effectively represents the texture at four different scales.

We now investigate how important it is to use four concentric circles to represent each pixel in both training and testing. The first row in Table 1 shows the performance when using four concentric circles with color to represent each pixel at the training stage, and four, two, and one circles also with color information for the testing data. The second row shows the performances when using the same number of circles to represent the pixels at the training and testing stage. It can be seen that performances in the first row are very similar so that four concentric circles is enough to represent the training data and fewer patches can be used to represent the pixels in the

TABLE 3

Performance Obtained for KNN and SVM Using pLSA or BoW Vectors for the Classifiers

	pLSA		BoW	
	KNN	SVM	KNN	SVM
C4CC	86.6	<b>87.1</b>	82.5	83.8
G4CC	84.3	84.7	79.7	80.8

OT database (eight categories) is used. Gray SIFT four Concentric Circles (G4CC) and Color SIFT with four Concentric Circles (C4CC).

testing images, that is, sampling only the training images at multiple scales is sufficient.

Table 2 summarizes the results for the three OT image sets (all eight categories, four natural and four man-made) covering the different dense vocabularies: gray and color patches, gray and color SIFT, and the two baseline algorithms when using KNN classifier. From these results, it can be seen that 1) The baseline texture algorithm works better than the baseline color in all three cases. Despite its simplicity, the performance of the baseline texture algorithm on man-made images (73.8 percent) is very high, showing that these images may be easily classified from their edge directions. 2) For the various descriptors, there are clear performance conclusions: man-made is always better classified than natural (as expected from the baseline results); SIFT-type descriptors are always superior to patches; color is always superior to gray level. The best performance (86.6 percent for all eight categories) is obtained using color SIFT and four concentric circles. 3) Somewhat surprisingly, better results are obtained using the SIFT vocabulary alone, rather than when merging both vocabularies (patches and SIFT). This may be because the parameters ( $V$ ,  $Z$ , and  $K$ ) have been optimized for a single vocabulary not under the conditions of using multiple vocabularies.

### 6.3 KNN versus SVM

All the results above are for  $P(z|d)$  with the KNN classifier. Now, we investigate classification performance when using an SVM. Table 3 shows the results for the SIFT support regions for both classifiers KNN and SVM. Optimized parameters for each vocabulary are used. It can be seen that SVM performs around 1 percent better than KNN.

### 6.4 pLSA versus Bag-of-Words (BoW)

The results to this point use pLSA to obtain an intermediate representation with  $P(z|d)$  as the inputs for the classifiers. We now compare to the performance obtained by classifying the BoW representation directly. Again, the performance is for the OT data set with eight categories, and in all the experiments:  $V = 1,500$  (unless stated otherwise),  $Z = 25$ ,  $K = 10$ , and four support regions are used for each point spaced at  $M = 10$ . For the SVM classifier, a  $\chi^2$  exponential kernel [36] is used for the BoW, and a euclidean exponential kernel for pLSA. These kernels were found to give the best performance in each case.

Table 3 shows pLSA and BoW rates for different support regions and using an SVM and a KNN. It can be seen that in all cases, the performance using pLSA is around 4 percent better than that obtained using a BoW.

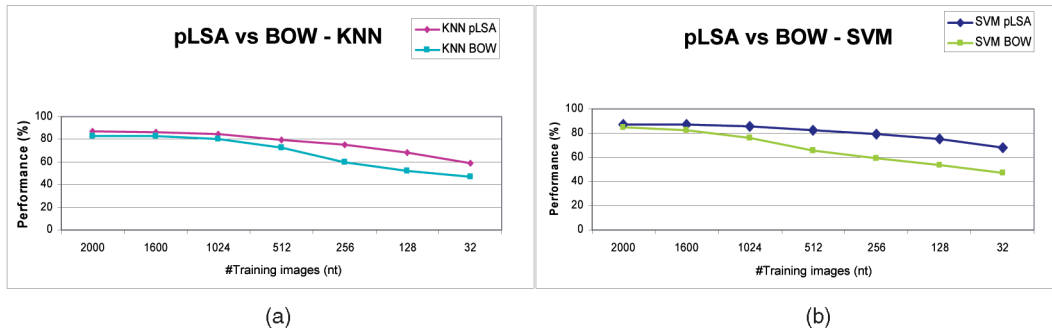


Fig. 6. pLSA and BoW performances when decreasing the number of training images. Eight categories from the OT data set with four concentric circles and  $V = 1,500$  words,  $Z = 25$ , and  $K = 10$ .

**Number of training images.** We now evaluate the classification performance when less training data is available. The OT data set is split into 2,000 training images and 688 test images. A varying number,  $n_{train}$ , of images from the training set are used for both learning the pLSA topics (generative part) and learning the topic distribution of each scene (discriminative part). The classification performance using  $P(z|d)$  is compared to that of using BoW vectors. As can be seen in Fig. 6, the gap between pLSA and BoW increases as the number of labeled training images decreases, as was demonstrated in [25].

In the previous experiment, we varied the amount of training data for both: the generative and discriminative learning. However, a key advantage of the hybrid approach is that the generative part of the model can be trained on large amounts of unlabeled data (hence, discovering the structure of the data) so that relatively few labeled examples are needed for high accuracy. To show this advantage, we repeat the previous experiment training the generative classifier using the 2,000 training images and decreasing the number of labeled training images ( $n_{train}$ ) only for the discriminative classifier. Fig. 7 shows the comparison of the previous experiment and the current experiment when using SVM as a discriminative classifier. It can be seen that much better results are obtained when decreasing only the number of labeled training data than when reducing the training data in both learning parts. Therefore, there is a clear advantage of using a hybrid approach: The system has acceptable performances with less labeled training data.

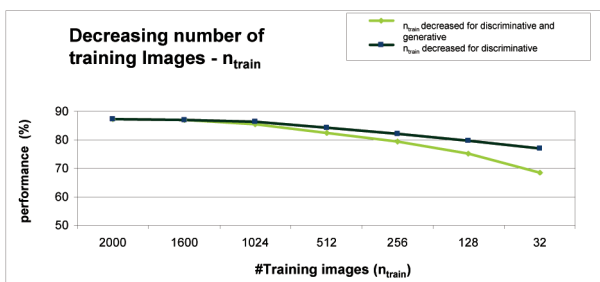


Fig. 7. Performance when decreasing the number of training images in the generative and discriminative parts (blue line) and when decreasing the number of labeled training images only for the discriminative part (yellow line). Eight categories from the OT data set with four concentric circles and  $V = 1,500$  words,  $Z = 25$ , and  $K = 10$ . An SVM is used as the discriminative classifier.

**Vocabulary Size.** Fig. 8 shows the performance when changing the vocabulary size  $V$  (from 200 to 5,000 words) for both the discriminative classifiers (KNN and SVM). It can be seen that for both classifiers, pLSA is less affected by the vocabulary size than the BoW.

**Number of scene categories.** Fig. 9 shows the performances when increasing the number of categories to be classified for both KNN (Fig. 9a) and SVM (Fig. 9b). For the KNN, when classifying the four natural images in the OT data set, the results using the topic distribution is 90.2 percent, and with the BoW directly, the classification performance decreases by only around 1.5 percent to 88.7 percent. However, for eight categories, the performance decreases by nearly 4 percent from 86.6 percent to 82.5 percent. Using the 13 categories from the FP data set and the 15 LSP data set, the performance falls to around 8 percent from 73.4 percent to 64.8 percent and from 71.0 percent to 63.1 percent, respectively. Thus, there is a clear gain in using pLSA (over the BoW) with KNN when classifying a large number of categories.

If we focus on the SVM, performances with pLSA are better as well. However, when classifying a large number of categories (13 or 15), pLSA is 1 percent better than BoW; thus, the gap is not as large as when using the KNN classifier. Table 4 summarizes the performances for KNN and SVM over pLSA and BoW.

## 6.5 Summary

The best results are obtained using dense descriptors—color SIFT with four circular support regions. Overlap increases the performance. When using the SIFT vocabulary, the values for the parameters giving the best results are  $M = 10$  pixels with concentric circles support regions of  $r = 4, 8, 12$ , and 16 pixels.

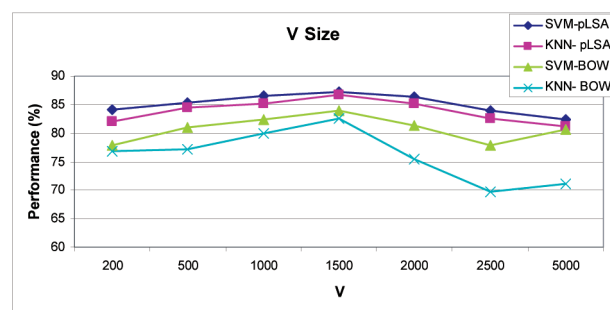


Fig. 8. Changing the vocabulary size for the OT data set. Parameters are  $Z = 25$ ,  $K = 10$ ,  $M = 10$ , and four concentric circles.



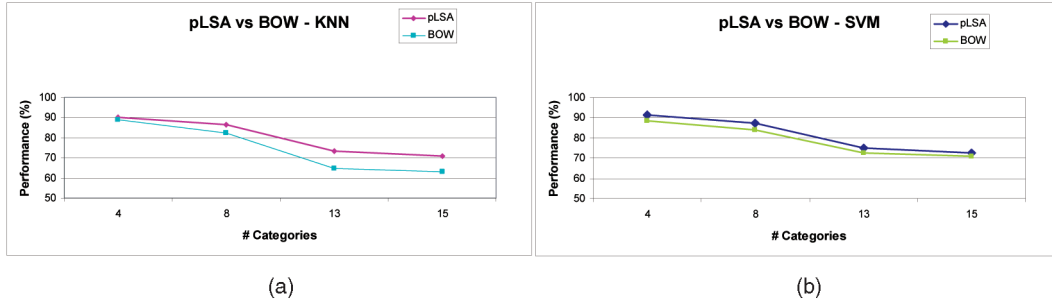


Fig. 9. pLSA and BoW performances when classifying different number of categories (from 4 to 15). Parameters used are  $V = 1,500$ ,  $Z = 25$ ,  $M = 10$ , and four concentric circles as support regions. (a) pLSA versus BoW when using KNN ( $K = 10$ ) and (b) pLSA versus BoW when using SVM.

For patches, the best results are for  $N = 5$  and  $M = 3$ . Table 5 shows the optimized values  $V$ ,  $Z$ , and  $K$  learned from a validation set for each data set. Note that  $V$  strongly depends on the size of the feature vector ( $128 \times 3$  dimensionality vector for SIFT, and  $25 \times 3$  dimensionality vector for patches), whereas  $Z$  depends on the number of categories in each data set. In both (SIFT and patches), color information increases performance. The result that dense SIFT gives the best performance was also found in [8] in the case of pedestrian detection. It is interesting that the same feature applies both to more distributed categories (like grass and mountains), as well as the compact objects (pedestrians) of their work where essentially only the boundaries are salient.

When comparing the discriminative classifiers KNN and SVM, better performances are obtained with SVM. We also demonstrated that pLSA works better than the BoW representation (pLSA provides a better intermediate representation of the images) and that pLSA is less affected by the vocabulary size and the number of training images. More concretely, for the KNN discriminative classifier, when working with a small number of categories, the difference between pLSA and BoW is 1.5 percent. However, when the number of categories increases, this difference is around 8 percent showing that pLSA provides a more robust intermediate representation than BoW. Thus, there is a clear gain in using pLSA (over the BoW) with KNN when classifying a large number of categories. Moreover, a clear advantage of using a generative model (pLSA), over BoW directly, is that the number of *labeled* training images can be reduced considerably without much loss of performance.

TABLE 4  
Classification Rates for pLSA and BoW when Classifying Categories from Different Data Sets

# Categ.	KNN		SVM	
	pLSA	BoW	pLSA	BoW
4 OT dataset	90.2	88.7	<b>91.5</b>	88.4
8 OT dataset	86.6	82.5	<b>87.1</b>	83.8
13 FP dataset	73.4	64.8	<b>74.9</b>	73.6
15 LSP dataset	71.0	63.1	<b>72.6</b>	72.5

Parameters used are  $V = 1,500$ ,  $Z = 25$ ,  $M = 10$ , and four concentric circles as support regions.

## 7 SPATIAL INFORMATION

Recently, it has been shown [2], [10], [18] that position information can improve scene classification performance (earlier work had shown little benefit [32]). Motivated by this, we add position information into our pLSA framework. We have implemented and compared four methods described below. For these results, the color SIFT vocabulary is used with four concentric circles spaced at  $M = 10$ , and the SVM is used as the discriminative classifier. The OT data set with optimized values (see Table 5) is used to evaluate performance.

**xy-pLSA.** The  $x$  and  $y$  normalized position of each pixel is concatenated to the feature vector. Therefore, now, the dimension of the feature vector is  $128 \times 3 + 2$ . Each component of the feature vector (both spatial and SIFT) is in the range  $[0, 1]$ . However, the SIFT part of the vector is sparse in general.

**ABS-pLSA.** This is the method proposed in [10]. The pLSA model is extended to incorporate location information by quantizing the location within the image into one of  $X$  bins. The joint density on the appearance and location of each region is then represented. Thus,  $P(w|z)$  in pLSA becomes  $P(w, x|z)$ , a discrete density of size  $(W \times X) \times Z$ . The same pLSA update equations outlined in Section 2 can be easily applied to this model in learning and recognition. The method is evaluated for  $X = 1, 4$ , and 16 bins, with the case  $X = 1$  corresponding to standard pLSA with no spatial information.

**Spatial pyramid matching (SPM).** This is the method proposed by Lazebnik et al. [18], which is based on spatial

TABLE 5  
Optimized Parameters when Using the SIFT Vocabulary for the Four Data Sets:  $M = 10$  and  $r = 4, 8, 12$ , and 16 Pixels, and when Using the Patch Vocabulary:  $N = 5$  and  $M = 3$  Pixels

Dataset	SIFT			Patch		
	$V$	$Z$	$K$	$V$	$Z$	$K$
VS	1500	25	7	900	25	9
OT	1500	25	10	900	23	10
FP	1200	35	9	600	33	10
LSP	1200	40	11	700	42	12

A validation set is used for each data set.

**TABLE 6**  
Performance Comparison for the OT Data Set when Spatial Information Is Used

Pyramid level	pLSA	xy-pLSA	ABS-pLSA	SPM	SP-pLSA
L = 0	87.1	89.0	87.1( $X = 1$ )	83.8	87.1
L = 1	—	—	87.9( $X = 4$ )	90.3	90.7
L = 2	—	—	88.3( $X = 16$ )	91.0	<b>91.1</b>

Four concentric circles spaced at  $M = 10$ , and  $V = 1,500$ ,  $Z = 25$ .

pyramid matching [13]. Pyramid matching works by placing a sequence of increasingly coarser grids over the feature space (in this case, over the image) and taking a weighted sum of the number of matches that occur at each level of resolution ( $L$ ). At any fixed resolution, two points are said to match if they fall into the same bin of the grid; matches found at finer resolutions are weighted more highly than matches found at coarser resolutions ( $\alpha_l$  represents the weight at level  $l$ ). The resulting spatial pyramid is an extension of the BoW image representation, it reduces to a standard BoW when  $L = 0$ , and a level 1 grid is equivalent to  $X = 4$  in the ABS-pLSA model.

**Spatial Pyramid—pLSA (SP-pLSA).** This method is inspired by both the previous ones, ABS-pLSA and SPM. We incorporate location information in pLSA by using the  $X$  bins at each resolution level  $L$ , weighting the bins for each level ( $\alpha_l$ ) as in SPM. Note that in ABS-pLSA, only the bins for one resolution level are used, and in SP-pLSA, we use the weighted bins for  $L$  resolutions. Therefore, for example, when  $L = 1$ , if using ABS-pLSA, we have  $X = 4$  bins, and if we use SP-pLSA, we have  $X = 5$  bins (one bin for  $L = 0$  and four bins for  $L = 1$ ). Thus,  $P(w|z)$  in pLSA becomes  $P(w, x, l|z)$ . The same pLSA update equations outlined in Section 2 can be easily applied to this model in learning and recognition.

Table 6 shows the values for pLSA without position and the four methods described above. The weights used in these experiments are  $\alpha_0 = 0.25$ ,  $\alpha_1 = 0.25$ , and  $\alpha_2 = 0.5$  (the same weights are used in [18]). When only the first level of the pyramid is used ( $L = 0$ ), the best result (89 percent) is obtained when using xy-pLSA. In this case, SPM works directly over the BoW and has worse results than the methods

**TABLE 7**  
Optimized Weight Ratios  $\alpha_0 : \alpha_2$  and  $\alpha_1 : \alpha_2$  for Each Data Set Using the Validation Set

ratios	OT - 8	OT - 4N	OT - 4MM	VS	FP	LSP
SPM weights						
$\alpha_0 : \alpha_2$	1	0.8	0.9	1	0.9	1
$\alpha_1 : \alpha_2$	0.9	0.9	0.8	0.9	0.8	0.8
SP-pLSA weights						
$\alpha_0 : \alpha_2$	0.7	0.9	0.9	0.9	1	1
$\alpha_1 : \alpha_2$	0.8	0.8	0.8	0.7	0.9	0.8

4N = 4 Natural categories; 4MM = 4 Man-Made categories.

that use pLSA. When  $L = 1$  and  $L = 2$ , the best results are obtained for SP-PLSA (90.7 percent and 91.1 percent) followed by SPM (90.3 percent and 91 percent). We only explored up to  $L = 2$ , which was demonstrated in [18] to be the optimum level.

**7.1 Weight Optimization— $\alpha_0$ ,  $\alpha_1$ , and  $\alpha_2$  on the Validation Set**

Using the validation set (see Section 5.2), we optimize the ratio between the weights  $\alpha_0 : \alpha_2$  and  $\alpha_1 : \alpha_2$  over the range [0, 1.5]. Fig. 10a shows the test performance when optimizing using SPM on the OT data set (eight categories). In this case, the best performance (92.2 percent for the test data) is for  $\alpha_0 : \alpha_2 = 1$  and  $\alpha_1 : \alpha_2 = 0.9$  (weights obtained from the validation set). This performance exceeds that given in Table 6.

Fig. 10b shows performances when optimizing the weights for SP-pLSA on the OT data set (eight categories). Now, the performance increases to 92.7 percent for the test data using the validation set optimized ratios  $\alpha_0 : \alpha_2 = 0.7$  and  $\alpha_1 : \alpha_2 = 0.8$ . Note that the best performances are obtained for higher ratios than in the experiments in Table 6 (where the ratio was  $\alpha_0 : \alpha_2 = 0.5$ , and  $\alpha_1 : \alpha_2 = 0.5$ ), which are the same ratios used in [18]. The optimized ratios using the validation set for all data sets used are summarized in Table 7. Default values of  $\alpha_0 : \alpha_2 = 0.9$  and  $\alpha_1 : \alpha_2 = 0.8$  clearly give superior performance than those in [18].

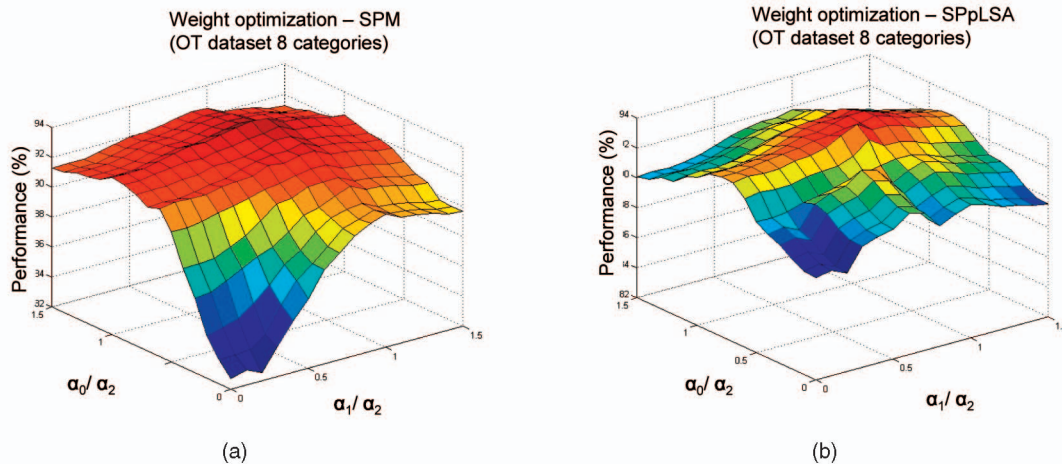


Fig. 10. Optimization rates between the weights at each pyramid level using the validation set from the OT data set: (a) SPM is used. (b) SP-pLSA is used.

TABLE 8  
Comparison of Our Algorithm with Other Methods Using Their Own Databases

Dataset	# of categ.	# train	# test	pLSA	SP-pLSA	SPM	Authors
OT	8	800	1888	82.5	<b>87.8</b>	87.1	83.7 [24]
OT	4 Natural	1000	472	90.7	<b>93.9</b>	93.3	89.0 [24]
OT	4 Man-Made	1000	216	91.7	<b>94.8</b>	94.2	89.0 [24]
VS	6	600	100	87.8	88.3	<b>88.6</b>	74.1 [32]
FP	13	1300	2459	74.3	<b>85.9</b>	85.5	65.2 [9]
LSP	15	1500	2986	72.7	<b>83.7</b>	83.5	81.4 [18]

Validation set optimized values are used for each data set.  $L = 2$  for SP-pLSA and SPM.

## 8 COMPARISON TO PREVIOUS RESULTS

### 8.1 Scene Classification

We compare the performance of our scene classification algorithm to the supervised approaches of Vogel and Schiele [32] and Oliva and Torralba [24], and the semisupervised approach of Fei-Fei and Perona [9] and Lazebnik et al. [18] using the same data sets that they tested their approaches on and the same number of training and testing images. For each data set, we use the SVM classifier, SIFT and four circular supports spaced at  $M = 10$ ; the parameters  $V$ ,  $Z$ ,  $\alpha_0$ ,  $\alpha_1$ , and  $\alpha_2$  have the validation set optimized values for each data set (see Table 5 and Table 7). We used color for OT and VS and gray for FP and LSP. The visual vocabulary is computed independently for each data set, as described in Section 5.2. We return to the issue of sharing vocabularies across data sets in Section 9. The results are given in Table 8.

Note that much better results are obtained with the four natural scenes of OT than with the six of VS. This is because the images in VS are much more ambiguous than those of OT and, consequently, more difficult to classify. Without using spatial information (fifth column in Table 8), our method outperforms the previous methods in [24], [9], [18], [32] despite the fact that our training is unsupervised in the sense that the scene identity of each image is unknown at the pLSA stage and is not required until the KNN or SVM training step. This is in contrast to that in [9], [18], where each image is labeled with the identity of the scene to which it belongs during the training stage. In [32], the training requires manual annotation of nine semantic concepts for 60,000 patches, whereas in [24], training requires manual annotation of six properties for thousands of scenes. It is worth noting that in [24], [32], the intermediate information that represents the images has a semantic meaning, whereas in [9], [18], and our approach, the intermediate information need not have a semantic meaning from the human point of view. However, this is not a problem: We are interested in the semantic meaning of the whole scene and not the intermediate information, because our final goal is to give a label for each scene.

As we noted in Section 7, better results are obtained with spatial information (sixth and seventh columns in Table 8). We have better performances than that in [18] when using SP-pLSA and also when using their own method with our features. Moreover, for a better comparison, we use the same number of words and weight ratios as in [18] ( $V = 200$  and  $L = 2$ ): they achieve 81.1 percent of correct classified scenes,

and we increase this to 82.2 percent (with SPM) when using four concentric circles to represent each pixel in the image. In both our and their experiments, gray SIFT descriptors are used. This demonstrates again that using more than one patch to represent each pixel increases performance.

#### 8.1.1 Discussion

The superior performance (compared to that in [9], [32]) could be due to the use of better features and how they are used. In the case of Vogel and Schiele [32], they learn nine topics (called *semantic concepts*) that correspond to those that humans can observe in the images: *water, trees, sky, etc.*, for six categories. In our case, we discover between 22 and 30 topics in the case of eight categories. These topics can vary depending if we are working with color features (where topics can distinguish objects with different colors like *light sky, blue sky, orange sky, orange foliage, green foliage, etc.*) or only gray SIFT features (objects like *trees and foliage, sea, buildings, and so forth*). In contrast to that in [32], we discover objects that sometimes would not be distinguished in a manual annotation, for example, *mountains with snow* and *mountains without snow*. Fei-Fei and Perona learn 40 topics (called *themes*) for 13 categories, but it is left unsaid whether these topics correspond to natural objects.

Our superior performance compared to that in [24] could be due to their method of scene interpretation. They propose a set of perceptual dimensions (for example, naturalness and openness) that represent the dominant spatial structure of a scene. These dimensions are estimated using spectral and coarsely localized information, using a very low-dimensional representation of the scene (Spatial Envelope), which bypasses the segmentation and the preprocessing of individual objects or regions. In contrast, in our approach, specific information about objects/topics is used for scene categorization. We also outperform the (SPM) classifier proposed in [18] when working with our pixel representation and features. Therefore, we have demonstrated that representing a pixel with more than one patch is better. Moreover, we successfully incorporated spatial information into the pLSA framework (SP-pLSA) obtaining slightly better performances than SPM though optimizing the level weights is responsible for the more significant part of the improvement.

### 8.2 Caltech 101

The Caltech-101 data set (collected by Fei-Fei et al. [20]) consists of images from 101 object categories and an additional background class, making the total number of

TABLE 9  
Classification of Caltech 101 with 15 or 30 Training Images per Class and 50 Test Images per Class

# train	SP-pLSA	SPM	[18]	[1]	[14]	[23]	[33]	[35]
15	<b>59.8</b> (±1.4)	58.7(±0.8)	56.4	52.0	49.5	51.9	44.0	59.0
30	<b>67.7</b> (±1.5)	66.5(±0.7)	64.6	—	58.2	56.0	63.0	66.0

For SP-pLSA and SPM, four concentric circles spaced at  $M = 10$  are used,  $V = 1,500$ ,  $Z = 80$ , and SVM is used as the discriminative classifier.

classes 102. This database contains from 31 to 800 images per category. Most images are medium resolution, about  $300 \times 300$  pixels. The significance of this database is its large interclass variability. A number of previously published papers have reported results on this data set: Lazebnik et al. [18], Berg et al. [1], Grauman and Darrell [14], Zhang et al. [35], and so forth.

For the experiments, dense color SIFT with four support regions are used to represent each pixel, spaced at  $M = 10$ ,  $V = 1,500$ , and  $Z = 80$  topics. The weight ratios are  $\alpha_0/\alpha_2 = 0.9$  and  $\alpha_1/\alpha_2 = 0.8$  for SP-pLSA and  $\alpha_0/\alpha_2 = 1$  and  $\alpha_1/\alpha_2 = 0.9$  for SPM. An SVM is used as the classifier. We carried out experiments using 15 and 30 random training images per category and 50 random testing images per class (disjoint from the training images). The mean recognition rate per class is used so that more populous (and easier) classes are not favored. This process is repeated 10 times, and the average correctness rate is reported. Table 9 shows our results and those reported by other authors. Our best performance is when using the SP-pLSA algorithm with a mean recognition rate of 59.8 percent with 15 training images per class and 67.7 percent with 30 training images per class. This outperforms the results reported by Zhang et al. [35], which to our knowledge are the best until now.

## 9 APPLICATIONS

We applied the pLSA-based classifier in four other situations. The first one is also a classification task, but combining the images of two different data sets, the second is a Relevant Feedback (RF) application, the third is scene retrieval for the film *Pretty Woman* [Marshall, 1990], and in the fourth, we apply pLSA for image segmentation. In all the following, the descriptor is a dense color SIFT with circular support and  $V = 700$ ,  $Z = 22$ , and  $K = 10$  (these are the optimal parameter values when working with the four natural scenes from the OT data set).

**Vocabulary generalization.** In this classification test, we train the system with the four natural scenes of the OT data set (*coast*, *forest*, *mountains*, and *open country*) and test using the same four scene categories from the VS data set. This tests whether the vocabulary and categories learned from one data set generalize to another. We obtain a performance of 88.2 percent of correctly classified images for KNN and 88.9 percent for SVM. This performance is only slightly worse than the 89.8 percent obtained when classifying the same four categories in the VS data set with no generalization (that is, using training images only from VS). This slight performance drop is because 1) images within the same database are more similar and 2) the images in VS are more ambiguous than OT, so this ambiguity is not represented in training the OT classifier. However, 88.9 percent compared to 89.8 percent does demonstrate excellent generalization. To address reason 1, we investigate using a

vocabulary composed from both databases and find that this improves the performance to 89.6 percent.

**RF.** Zhang and Zhang [37] proposed a method for improving the retrieval performance, given a probabilistic model. It is based on moving the query point in the visual word space toward good example points (relevant images) and away from bad example points (irrelevant images). The vector moving strategy uses the Rocchio's formula [26]

$$q_{pos} = \alpha q + \beta \left( \frac{1}{a} \sum_{i=1}^a rel_i \right) - \gamma \left( \frac{1}{b} \sum_{j=1}^b irel_j \right), \quad (3)$$

where  $q$  is the BoW for the query image,  $a$  is the number of relevant images,  $b$  is the number of irrelevant images, and  $rel$  and  $irel$  are the BoW representations for the relevant and irrelevant retrieved images. The parameters  $\alpha$ ,  $\beta$ , and  $\gamma$  are set to 1. With the modified query vector  $q_{pos}$  and a constructed negative example  $q_{neg}$

$$q_{neg} = \alpha \left( \sum_{j=1}^b irel_j \right) + \beta \left( \frac{1}{b} \sum_{j=1}^b irel_j \right) - \gamma \left( \frac{1}{a} \sum_{i=1}^a rel_i \right). \quad (4)$$

Their representations in the discovered concept space are obtained by  $P(z|q_{pos})$  and  $P(z|q_{neg})$ , and their similarities  $sp_i$  and  $sn_i$  to each image  $i \in I$  in the database are measured using the cosine metric of the corresponding vectors in the topic space, respectively. Then, the images are ranked based on the similarity  $s_i = sp_i - sn_i$ .

To test RF, we simulate the user's feedback using 25 random images of each category. For each query image, we carry out  $n$  iterations. At each iteration, the system examines the top 20, 40, or 60 images that are most similar to the query excluding the positive examples labeled in previous iterations. Images from the same category as the initial query will be used as positive examples and other images as negative examples. We used 200 query images, 25 of each category, in the OT data set. Best results are obtained when considering the top 60 images. The first 100 images can be retrieved with an average precision of 0.75. The most difficult category to retrieve is *open country*, whereas the better retrieved are *forest* and *highway* followed by *tall buildings*. This is in accordance with the classification results.

**Classifying film frames into scenes.** In this test, the images in OT are again used as training images (eight categories), and keyframes from the movie *Pretty Woman* are used as test images. We used  $V = 1,500$  and  $Z = 25$ , which are the optimized values for the eight categories in the OT data set. Note that this is a second example of vocabulary and topic generalization as we are using training images from a different data set. We used every 100th frame from the movie to form the test set. In this movie, there are only a few images that could be classified as the same categories used in



Fig. 11. Example frames from the film *Pretty Woman* with their classification. The classifier is trained on the OT data set.

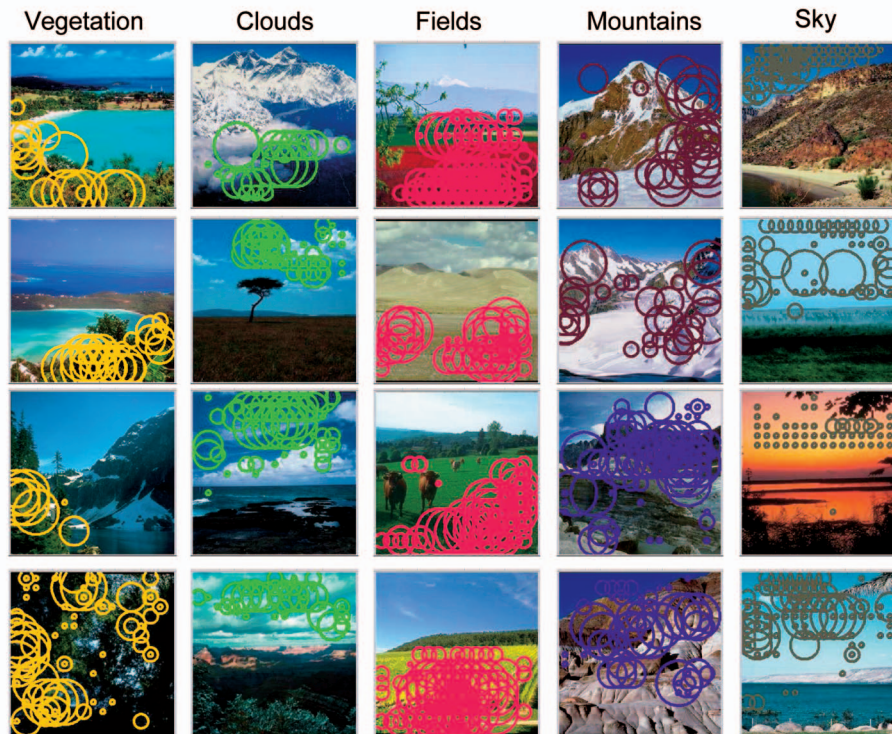


Fig. 12. Topics segmentation. Five topics (vegetation, clouds, fields, mountains, and sky) are shown. Only circular regions with a topic posterior  $P(z|w, d)$  greater than 0.8 are shown.

OT, and there are many images containing only people. Therefore, it is a difficult task for the system to correctly classify the keyframes. Although the results obtained (see Fig. 11) are purely anecdotal, they are very encouraging and show again the success of using pLSA in order to classify scenes according to their topic distribution.

**Segmentation.** Fig. 12 shows examples of segmentation of five topics using the color SIFT vocabulary. Circular patches are painted according to the maximum posterior  $P(z|w, d)$

$$P(z|w, d) = \frac{P(w|z)P(z|d)}{\sum_{z_1 \in Z} P(w|z_1)P(z_1|d)}. \quad (5)$$

For each visual word in the image, we choose the topic with maximum posterior  $P(z|w, d)$  and paint the patch with its associated color, so each color represents a different topic (the topic color is chosen randomly). To simplify the figures, we only paint one topic each time. Note that topics represent

consistent regions across images (enabling a coarse segmentation), and there is a straightforward correspondence between topic and object.

## 10 DISCUSSION—THE SCENE CLASSIFICATION TASK

Fig. 13a shows the confusion matrix between the eight categories in the OT data set when no spatial information is used. The best classified scenes are *highway* and *forest* with a performance of 89.8 percent and 98.8 percent, respectively. The most difficult scenes to classify are *open country*. There is confusion between the *open country* and *coast* scenes and between the *open country* and *mountain* scenes. The most confused man-made images are *street*, *inside city*, and *highway*. These are also the most confused categories in [24]. We can also establish some relationship among the categories by looking at the distances among the topic distributions between them (see the dendrogram in Fig. 13b). When the topic distributions are close, the categories are also close to

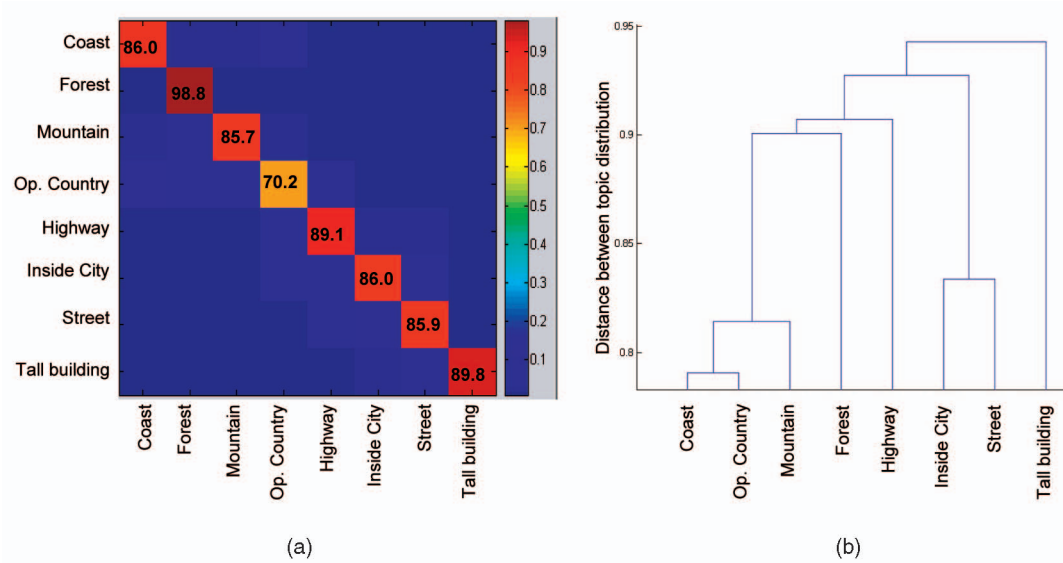


Fig. 13. (a) Confusion matrix for the eight categories in the OT data set using no spatial information. (b) Dendrogram showing the closest categories, which are also the most confused.

each other on the dendrogram. For example, the closest in the natural categories are *open country* and *coast*, and the closest in the man-made categories are *inside city* and *street*.

Fig. 14 shows some images confused between categories showing the ambiguity between some of them. Scene categorization is characterized by potential ambiguities since it depends strongly on the subjective perception of the viewer. For example, some of the *open country* images shown in Fig. 14a can be easily classified as *mountain* for some humans as the system did. Obviously, the obtainable classification accuracies depend strongly on the consistency and accuracy of the manual annotations, and sometimes annotation ambiguities are unavoidable. For example, the annotation of *mountains* and *open country* is quite challenging.

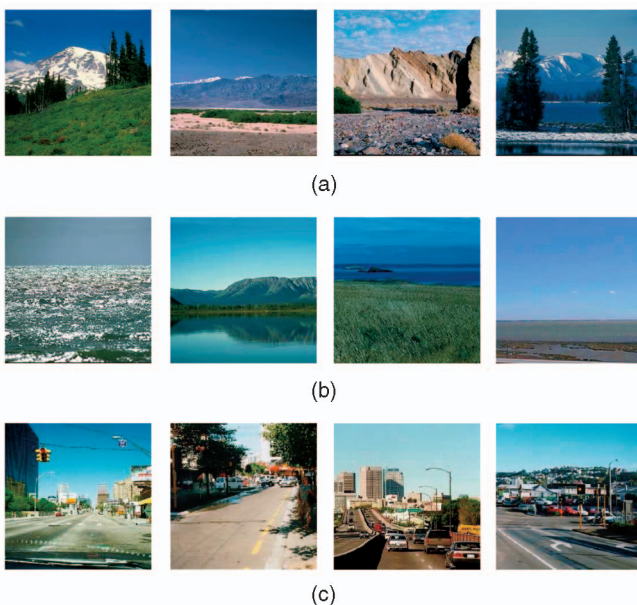


Fig. 14. Images showing the most confused categories: (a) open country images classified as mountains, (b) coast images classified as open country, and (c) highway images classified as street.

Imagine an image with *fields* and *snow hills* in the far distance: Is it *open country* or *mountain*? Even more confused are *coast* and *open country* scenes (Fig. 14b), yet both of them have a similar structure: *water* or *fields* and the *sky* in the distance. For that reason, it is not surprising that *coast* and *open country* are confused in both directions. Another major confusion appears between *streets* and *highway*. This results mainly from the fact that each street scene contains a *road*, whereas the most important part of highway scenes is the *road*. *Streets* and *inside city* images are confused because, normally, streets occur in cities.

Fig. 15a shows the confusion matrix for the eight categories in the OT data set when using SP-pLSA with  $L = 2$ . Now, for the *forest* scenes, we obtain a rate of 100 percent of correct classified images, and all the classification rates for the other scenes are also increased. Again, the most difficult scenes to classify are the *open country*. Fig. 15b shows some images well classified using SP-pLSA and poorly classified without spatial information. This demonstrates that spatial distribution can reduce the ambiguity—or at least that spatial distribution correlates with the annotator’s choices. However, we are still far from 100 percent correct classification, again, due to the ambiguities between the scene categories used. Vogel and Schiele [32] analyzed in detail the ambiguities between scene categories, showing that there is a semantic transition between categories. Their experiments with human subjects showed that many images cannot be clearly assigned to one category. How far away must a *mountain* be so that the image moves from the *mountains* category to the *open country* category? How much *road* is necessary to make a *street* image into a *highway* image, and vice versa? In addition, we arrive at the same conclusion as that in [32]: It is not wise to aim for a hard decision categorization of scenes. However, since scenes, that is full images, contain very complex semantic details, hard scene categorization is an appropriate task for 1) testing the image representation [32], in this case, provided by topics, 2) as an approximation for how the ranking on an image retrieval system would work, and 3) classifying mutually exclusive scenes such as indoor/ outdoor, garden/bathroom, or coast/kitchen.

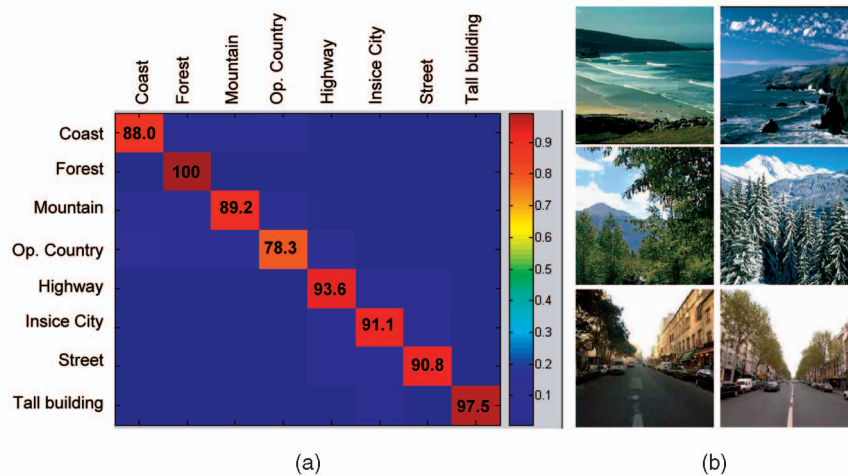


Fig. 15. (a) Confusion matrix for the eight categories in the OT data set when using SP-pLSA. (b) Top: Two coast scenes classified as mountain when spatial information is not used, and correctly classified using SP-pLSA. Middle: two forest scenes classified as mountains without spatial information, and correctly classified with SP-pLSA. Bottom: Two street scenes classified as highway without spatial information and correctly classified using SP-pLSA.

We have done some preliminary experiments with k-means clustering the image topics provided by SP-pLSA to automatically detect visually similar categories. The results are interesting because the resulting clusters have a semantic meaning such as fields with mountains at the back, fields with flowers, coasts with rocks, sunshine coast, highway with cars and without cars, and so forth. Nevertheless, the images with a semantic transition between categories are not well clustered (because there are not sufficient ambiguous images). A solution would be to use EM soft assignment in the clustering.

## 11 CONCLUSIONS

We have proposed a scene classifier that learns topics and their distributions in unlabeled training images using pLSA, and then uses their distribution in test images as a feature vector in a supervised discriminative classifier. In contrast to previous approaches [9], [24], [32], our topic learning stage is completely unsupervised, and we obtain significantly superior performance in the pure bag of words situation (no spatial information). We also have shown that the pLSA adapted to incorporate spatial information at different resolution levels (SP-pLSA) has comparable/slightly superior performance with the spatial pyramid matching proposed in [18]. We also demonstrated that using more than one patch to represent each pixel gives better performance, outperforming the method in [18] when using their own approach with spatial information.

We studied the influence of various descriptor parameters and have shown that using dense SIFT descriptors with overlapping patches gives the best results for man made, as well as for natural scene classification. Furthermore, discovered topics correspond fairly well with different textural objects (grass, mountains, and sky) in the images, and topic distributions are consistent between images of the same category. It is probably this freedom in choosing appropriate topics for a data set, together with the optimized features and vocabularies, that is responsible for the superior performance of the scene classifier over previous nonspatial work (even in cases where manual annotation was provided). Moreover,

the use of pLSA is never detrimental to performance, and it gives a significant improvement over the original BoW model when a large number of scene categories are used.

## ACKNOWLEDGMENTS

The authors would like to thank Antonio Torralba, Julia Vogel, Fei-Fei Li, and Lazebnik Lazebnik for providing their data sets and Ondra Chum, Rob Fergus, Jan-Mark Geusebroek, David Lowe, Cordelia Schmid, Josef Sivic, Bill Triggs, and Tinne Tuytelaars for discussions. This work was partially funded by the research grant BR03/01 from the University of Girona, the EU NOE PASCAL, and the EU Project CLASS.

## REFERENCES

- [1] A.C. Berg, "Shape Matching and Object Recognition," PhD dissertation, Computer Science Division, Univ. of California, 2005.
- [2] A. Bosch, X. Muñoz, and J. Freixenet, "Segmentation and Description of Natural Outdoor Scenes," *Image and Vision Computing*, vol. 25, no. 5, pp. 727-740, May 2006.
- [3] A. Bosch, X. Muñoz, and R. Martí, "A Review: Which Is the Best Way to Organize/Classify Images by Content," *Image and Vision Computing*, vol. 25, no. 6, pp. 778-791, June 2007.
- [4] A. Bosch, A. Zisserman, and X. Muñoz, "Scene Classification via pLSA," *Proc. European Conf. Computer Vision*, vol. 4, pp. 517-530, May 2006.
- [5] C. Chang and C. Lin, *LIBSVM: A Library for Support Vector Machines*, 2001.
- [6] J.L. Crowley and F. Berard, "Multi-Modal Tracking of Faces for Video Communication," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, pp. 640-645, 1997.
- [7] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual Categorization with Bags of Keypoints," *Proc. Workshop Statistical Learning in Computer Vision*, pp. 1-22, May 2004.
- [8] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 886-893, June 2005.
- [9] L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, pp. 524-531, 2005.
- [10] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning Object Categories from Google's Image Search," *Proc. Int'l Conf. Computer Vision*, vol. II, pp. 1816-1823, Oct. 2005.

- [11] G.D. Finlayson, B. Schiele, and J.L. Crowley, "Comprehensive Colour Normalization," *Proc. European Conf. Computer Vision*, vol. 1, pp. 475-490, 1998.
- [12] T. Geodeme, T. Tuytelaars, G. Vanacker, M. Nuttin, and L. Van Gool, "Omnidirectional Sparse Visual Path Following with Occlusion-Robust Feature Tracking," *Proc. OMNIVIS Workshop, Int'l Conf. Computer Vision*, vol. 3115, pp. 207-215, Oct. 2005.
- [13] K. Grauman and T. Darrell, "The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features," *Proc. Int'l Conf. Computer Vision*, pp. 1458-1465, 2005.
- [14] K. Grauman and T. Darrell, "Pyramid Match Kernels: Discriminative Classification with Sets of Image Features (Version 2)," Technical Report CSAIL-TR-2006-020, Computer Science and Artificial Intelligence Laboratory, Massachusetts Inst. of Technology, 2006.
- [15] T. Hofmann, "Probabilistic Latent Semantic Indexing," *Proc. SIGIR Conf. Research and Development in Information Retrieval*, 1998.
- [16] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, vol. 41, no. 2, pp. 177-196, 2001.
- [17] S. Lazebnik, C. Schmid, and J. Ponce, "A Sparse Texture Representation Using Affine-Invariant Regions," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 319-324, June 2003.
- [18] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 2169-2178, June 2006.
- [19] T. Leung and J. Malik, "Representing and Recognizing the Visual Appearance of Materials Using Three-Dimensional Textons," *Int'l J. Computer Vision*, vol. 43, no. 1, pp. 29-44, June 2001.
- [20] F. Li, R. Fergus, and P. Perona, "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories," *Proc. Workshop Generative-Model Based Vision*, p. 178, June 2004.
- [21] D. Lowe, "Distinctive Image Features from Scale Invariant Keypoints," *Int'l J. Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
- [22] K. Mikolajczyk and C. Schmid, "Scale and Affine Invariant Interest Point Detectors," *Int'l J. Computer Vision*, vol. 60, no. 1, pp. 63-86, 2004.
- [23] J. Mutch and D. Lowe, "Multiclass Object Recognition Using Sparse, Localized Features," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 11-18, June 2006.
- [24] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," *Int'l J. Computer Vision*, vol. 42, no. 3, pp. 145-175, 2001.
- [25] P. Quelhas, F. Monay, J.M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool, "Modeling Scenes with Local Descriptors and Latent Aspects," *Proc. Int'l Conf. Computer Vision*, pp. 883-890, Oct. 2005.
- [26] J.J. Rocchio, "Relevance Feedback in Information Retrieval," *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice Hall, 1971.
- [27] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman, "Discovering Objects and Their Locations in Images," *Proc. Int'l Conf. Computer Vision*, pp. 370-377, Oct. 2005.
- [28] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," *Proc. Int'l Conf. Computer Vision*, vol. 2, pp. 1470-1477, Oct. 2003.
- [29] M. Szummer and R.W. Picard, "Indoor-Outdoor Image Classification," *Proc. ICCV Workshop Content-Based Access of Image and Video Databases*, pp. 42-50, 1998.
- [30] A. Vailaya, A. Figueiredo, A. Jain, and H. Zhang, "Image Classification for Content-Based Indexing," *IEEE Trans. Image Processing*, vol. 10, pp. 117-129, 2001.
- [31] M. Varma and A. Zisserman, "Texture Classification: Are Filter Banks Necessary?" *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 691-698, June 2003.
- [32] J. Vogel and B. Schiele, "Semantic Modeling of Natural Scenes for Content-Based Image Retrieval," *Int'l J. Computer Vision*, vol. 72, no. 2, pp. 133-157, Jan. 2007.
- [33] G. Wang, Y. Zhang, and L. Fei-Fei, "Using Dependent Regions for Object Categorization in a Generative Framework," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 1597-1604, June 2006.
- [34] J. Weijer and C. Schmid, "Coloring Local Feature Extraction," *Proc. European Conf. Computer Vision*, vol. 2, pp. 332-348, May 2006.
- [35] H. Zhang, A. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 2126-2136, June 2006.
- [36] J. Zhang, M. Marszałek, C. Lazebnik, and S. Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study," *Int'l J. Computer Vision*, 2007.
- [37] R. Zhang and Z. Zhang, "Hidden Semantic Concept Discovery in Region Based Image Retrieval," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 996-1001, June 2004.



**Anna Bosch** received the engineering and master's degrees in computer science from the University of Girona, Spain, in 2003. She is currently working toward the PhD degree in the Computer Vision and Robotics Group, University of Girona. She has been working at the Visual Geometry Group at the University of Oxford as a visiting student since 2005. Her research interests include object and scene classification, machine learning, content-based image retrieval systems, and medical image applications.

**Andrew Zisserman** bio and photo were not available at press time.



**Xavier Muñoz** received the MS and PhD degrees in computer science from the University of Girona in 1999 and 2003, respectively. After completing the MS degree, he joined the Computer Vision and Robotics Group. In 2004, he founded the spin-off company DSET, where he works as a manager. At present, he is also an associate professor in the Electronics, Computer Engineering, and Automation Department of the University of Girona. His research interests include texture analysis, image segmentation, and object and image classification.

► For more information on this or any other computing topic, please visit our Digital Library at [www.computer.org/publications/dlib](http://www.computer.org/publications/dlib).