Scene Classification via pLSA

Anna Bosch[1], Andrew Zisserman[2], and Xavier Muñoz[1]

[1] Computer Vision and Robotics Group, University of Girona, 17071 Girona
{aboschr,xmunoz}@eia.udg.es
[2] Robotics Research Group, University of Oxford, Oxford OX1 3PJ
az@robots.ox.ac.uk

Abstract. Given a set of images of scenes containing multiple object categories (e.g. grass, roads, buildings) our objective is to discover these objects in each image in an unsupervised manner, and to use this object distribution to perform scene classification. We achieve this discovery 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. The scene classification on the object distribution is carried out by a k-nearest neighbour classifier.
We investigate the classification performance under changes in the visual vocabulary and number of latent topics learnt, and develop a novel vocabulary using colour SIFT descriptors. Classification performance is compared to the supervised approaches of Vogel & Schiele [19] and Oliva & Torralba [11], and the semi-supervised approach of Fei Fei & Perona [3] using their own datasets and testing protocols. In all cases the combination of (unsupervised) pLSA followed by (supervised) nearest neighbour classification achieves superior results. We show applications of this method to image retrieval with relevance feedback and to scene classification in videos.

1 Introduction

Classifying scenes (such as mountains, forests, offices) is not an easy task owing to their variability, ambiguity, and the wide range of illumination and scale conditions that may apply. Two basic strategies can be found in the literature. The first uses low-level features such as colour, texture, power spectrum, etc. This approaches consider the scene as an individual object [16, 17] and is normally used to classify only a small number of scene categories (indoor versus outdoor, city versus landscape etc...). The second strategy uses an intermediate representations before classifying scenes [3, 11, 19], and has been applied to cases where there are a larger number of scene categories (up to 13).

In this paper we introduce a new classification algorithm based on a combination of unsupervised probabilistic Latent Semantic Analysis (pLSA) [6] followed by a nearest neighbour 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". 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 (colour, texture, orientation). The approach is inspired in particular by three previous papers: (i) the use of pLSA on sparse features for recognizing compact object categories (such as Caltech cars and faces) in Sivic *et al.* [15]; (ii) the dense SIFT [9] features developed in Dalal and Triggs [2] for pedestrian detection; and (iii) the semi-supervised application of Latent Dirichlet Analysis (LDA) for scene classification in Fei Fei and Perona [3]. 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.* [12] 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 three previous methods [3, 11, 19] using the authors' own databases. The previous works used varying levels of supervision in training (compared to the unsupervised object discovery developed in this paper): Fei Fei and Perona [3] requires the category of each scene to be specified during learning (in order to discover the *themes* of each category); Oliva and Torralba [11] require a manual ranking of the training images into 6 different properties; and Vogel and Schiele [19] require manual classification of 59582 local patches from the training images into one of 9 *semantic concepts*. As will be seen, we achieve superior performance in all cases.

We briefly give an overview of the pLSA model in Section 2. Then in Section 3 we describe the classification algorithm based on applying pLSA to images. Section 4 describes the features used to form the visual vocabulary and the principal parameters that are investigated. A description of datasets and a detailed description of the experimental evaluation is given in Sections 5 and 6.

2 pLSA model

Probabilistic Latent Semantic Analysis (pLSA) is a generative model from the statistical text literature [6]. 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 *topics as object categories* (e.g. grass, houses), so that an image containing instances of several objects is modelled as a mixture of topics. The models are applied to images by using a *visual* analogue of a *word*, formed by vector quantizing colour, texture and SIFT feature like region descriptors (as described in Section 4). pLSA is appropriate here because it provides the 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 $D = d_1,...,d_N$ with words from a visual vocabulary $W = w_1,...,w_V$. One may summarize the data in 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 word w_i occurred in an image d_j. In pLSA there is also a latent variable

model for co-occurrence data which associates an unobserved class variable $z \epsilon Z = z_1,...,z_Z$ with each observation. A joint probability model $P(w,d)$ over $V \times N$ is defined by the mixture:

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

$P(w|z)$ are the topic specific distributions and, each image is modelled as a mixture of topics, $P(z|d)$. For a fuller explanation of the model refer to [5, 6, 15].

3 Classification

In training the topic specific distributions $P(w|z)$ are learnt from the set of training images. Each training image is then represented by a Z-vector $P(z|d_{train})$, where Z is the number of topics learnt. 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 (i.e. which category they are in) or any region segmentation.

Classification of an unseen test image proceeds in two stages. First the document specific mixing coefficients $P(z|d_{test})$ are computed, and then these are used to classify the test images using a K nearest neighbour scheme. In more detail document specific mixing coefficients $P(z|d_{test})$ are computed using the fold-in heuristic described in [5]. The unseen image is projected onto the simplex spanned by the $P(w|z)$ learnt during training, i.e. the mixing coefficients $P(z_k|d_{test})$ are sought such that the Kullback-Leibler divergence between the measured empirical distribution and $P(w|d_{test}) = \sum_{z \epsilon Z} P(w|z)P(z|d_{test})$ 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 learnt $P(w|z)$ kept fixed. The result is that the test image is represented by a Z-vector. The test image is then classified using a K Nearest Neighbours classifier (KNN) on the Z-vectors of the training images. An Euclidean distance function is used. In more detail, the KNN selects the K nearest neighbours of the new image within the training database. Then it assigns to the new picture the label of the category which is most represented within the K nearest neighbours. Figure 1 shows graphically the learning and classification process.

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 in the first part of Figure 1. Previously both sparse [1, 7, 14] and dense descriptors, e.g. [2, 8, 18], 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.

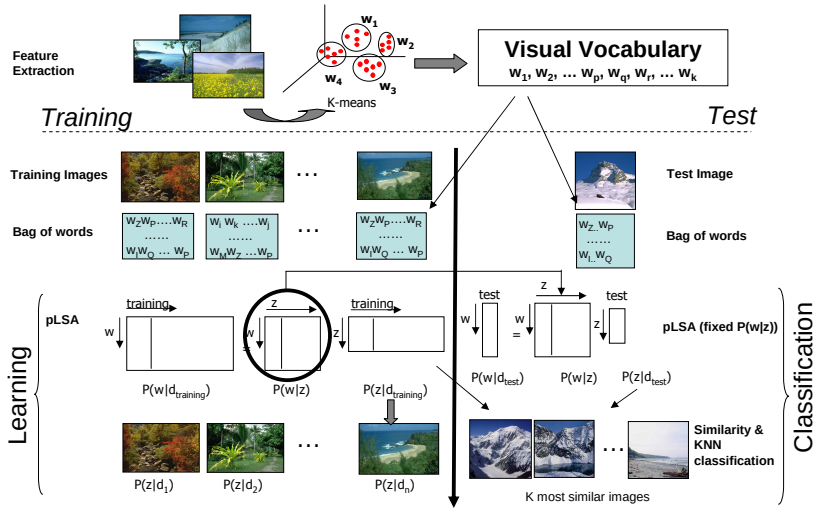


Fig. 1. Overview of visual vocabulary formation, learning and classification stages.

We vary the normalization, sizes of the patches, and degree of overlap. The words produced are evaluated by assessing their classification performance over three different databases 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:

Grey patches (dense). As in [18], and using only the grey level information, the descriptor is a N $\times$ N square neighbourhood 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 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 $N = 11$).

Colour patches (dense). As above, but the colour information is used for each pixel. We consider the three colour components HSV and obtain a $N^2 \times 3$ dimensional vector.

Grey SIFT (dense). SIFT descriptors [9] 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/or 16 pixels. Consequently each point is represented by n SIFT descriptors (where n is the number of circular supports), each is 128-dim. When $n > 1$, 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.

Colour SIFT (dense). As above, but now SIFT descriptors are computed for each HSV component. This gives a 128 $\times$ 3 dim-SIFT descriptor for each point.

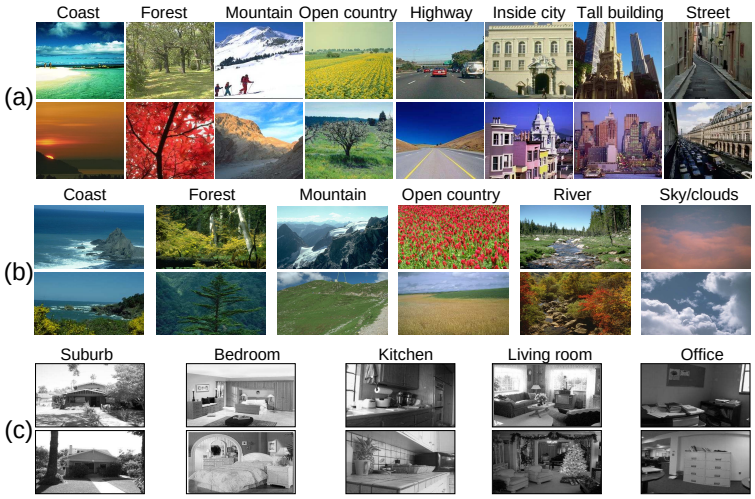


Fig. 2. Example images from the three different datasets used. (a) from dataset OT [11], (b) from dataset VS [19], and (c) from the dataset FP [3]. The remaining images of this dataset are the same as in OT but in greyscale.

Note, this is a novel feature descriptor. Another way of using colour with SIFT features has been proposed by [4].

Grey SIFT (sparse). Affine co-variant regions are computed for each grey scale image, constructed by elliptical shape adaptation about an interest point [10]. These regions are represented by ellipses. Each ellipse is mapped to a circle by appropriate scaling along its principal axis and a 128-dim SIFT descriptor computed. This is the method used by [1, 7, 14, 15].

5 Datasets and Methodology

5.1 Datasets

We evaluated our classification algorithm on three different datasets: (i) Oliva and Torralba [11], (ii) Vogel and Schiele [19], and (iii) Fei Fei and Perona [3]. We will refer to these datasets as OT, VS and FP respectively. Figure 2 shows example images from each dataset, and the contents are summarized here:

OT: includes 2688 images classified as 8 categories: 360 coasts, 328 forest, 374 mountain, 410 open country, 260 highway, 308 inside of cities, 356 tall buildings, 292 streets. The average size of each image is 250×250 pixels.

VS: includes 702 natural scenes consisting of 6 categories: 144 coasts, 103 forests, 179 mountains, 131 open country, 111 river and 34 sky/clouds. The size of the images is 720×480 (landscape format) or 480×720 (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 greyscale. This dataset consists of the 2688 images (8 categories) of the OT dataset plus: 241 suburb residence, 174 bedroom, 151 kitchen, 289 living room and 216 office. The average size of each image is approximately 250×300 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 overall performance rates are measured by the average value of the diagonal entries of the confusion table.

Datasets are split randomly into two separate sets of images, half for training and half for testing. We take 100 random images from the training set 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 learnt from about 30 random training images of each category.

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:

Global colour model. The algorithm computes global HSV histograms for each training image. The colour values are represented by a histogram with 36 bins for H, 32 bins for S, and 16 bins for V, giving a 84-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 training image (greyscale). These orientations are collected into a 72 bin histogram for each image. The classification of a test image is again carried out using KNN.

Moreover the KNN classifier is also applied directly to the bag-of-words (BOW) representation (i.e. to $P(w|d)$) in order to assess the gain in using pLSA (where the KNN classifier is applied to the topic distribution $P(z|d)$).

6 Classification results

We investigate the variation of classification performance with change in visual vocabulary, number of topics etc for the case of the OT dataset. The results for the datasets FP and VS use the optimum parameters selected for OT and are given in Section 6.2 below. For the OT dataset three classification situations are considered: classification into 8 categories, and also classification within the two subsets of natural (4 categories), and man-made (4 categories) images. The latter two are the situations considered in [11]. We carry out experiments with normalized images (zero mean and unit standard deviation) and unnormalized images.

Excluding the preprocessing time of feature detection and visual vocabulary formation, it takes about 15 mins to fit the pLSA model to 1600 images (Matlab implementation on a 1.7GHz Computer).

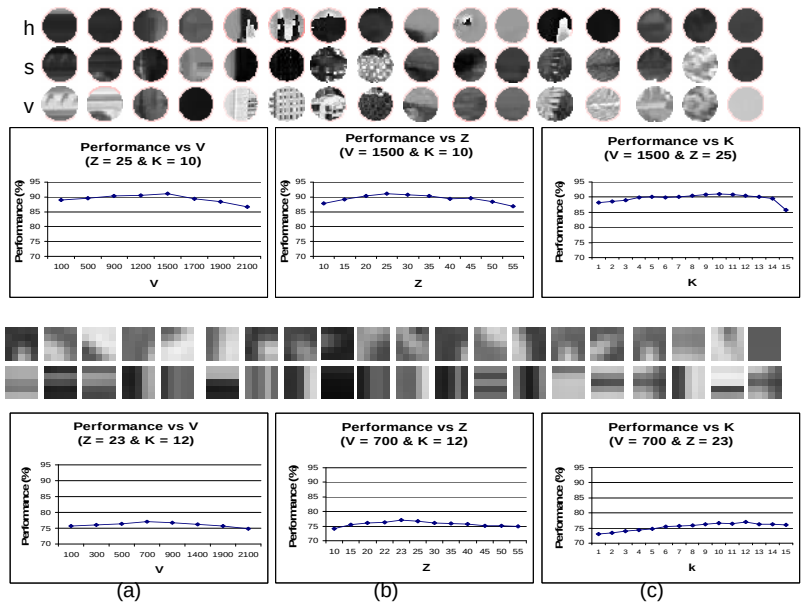
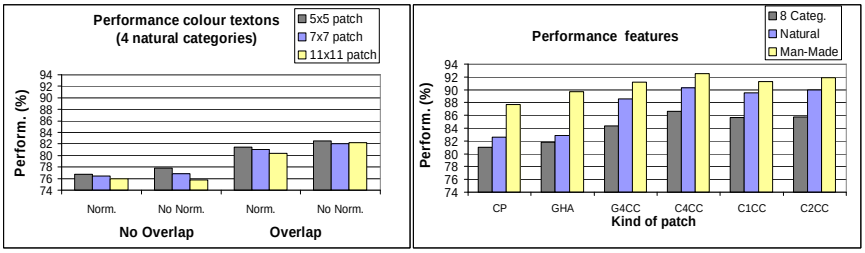


Fig. 3. Performance under variation in various parameters for the 8 category OT classification. Top: example visual words and performance for dense colour SIFT $M = 10$, $r = 4$, 8, 12 and 16 (each column shows the HSV components of the same word). Lower example visual words and performance for grey patches with $N = 5$ and $M = 3$. (a) Varying number of visual words, V, (b) Varying number of topics, Z, (c) Varying number k (KNN).

6.1 Classification of the OT dataset

We first investigate how classification performance (on the validation set – see Section 5.2) 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 neighbours (K in kNN). Figure 3 shows this performance variation for two types of descriptor – dense colour SIFT with $M = 10$ and four circular supports, and grey 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 colour SIFT the mode is at $V = 1500$ and $Z = 25$, while for grey patches the mode is at $V = 700$ and $Z = 23$. For K the performance increases progressively until K is between 10 and 12, and then drops off slightly. In the following results the optimum choice of parameters is used for each descriptor type.

To investigate the statistical variation we repeat the dense colour SIFT experiment ($r = 4$, 8, 12, 16 and $M = 10$) 15 times with varying random selection of the training and test sets, and building the visual vocabulary afresh each time. All parameters are fixed with the number of visual words $V = 1500$, the number of topics $Z = 25$ and the number of neighbours $K = 10$. We obtained



(a) (b)

Fig. 4. (a) The performance when classifying the four natural categories using normalized and unnormalized images and with overlapping and non-overlapping patches. Colour patches are used. (b) Performance when classifying all categories, man-made and natural using different patches and features. (CP = Colour patches - dense; GHA = Grey Harris Affine - sparse; G4CC = Grey SIFT concentric circles - dense; C4CC = Colour SIFT 4 concentric circles - dense; C1CC = Colour SIFT 1 Circle - dense; C2CC = Colour SIFT 2 concentric circles - dense.

performance values between 79% and 86% with a mean of 84.78% and standard deviation of 1.93%.

We next investigate the patch descriptors in more detail. Figure 4a shows the results when classifying the images of natural scenes with colour-patches. The performance when using unnormalized images is nearly 1% better than when using normalized. When using overlapping patches, the performance increases by almost 6% compared to no overlap. Similar results occur for the man-made and all scene category sets. Comparing results when classifying the images using only grey level information or using colour, it can be seen that colour brings an increment of around 6-8%. This is probably because colour is such an important factor in outdoor images, and helps to disambiguate and classify the different objects in the scene. For colour patches the best performance is obtained when using the 5×5 patch over unnormalized images, with $M = 3$, $V = 900$, $Z = 23$ and $K = 10$.

The performance of SIFT features is shown in Figure 4b. 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 colour is a benefit with better results obtained using colour than grey SIFT. The performance using grey SIFT when classifying natural images is 88.56% and increase 2% when using colour SIFT, both with four concentric support regions. The difference when using these vocabularies with man-made images is not as significant. This reiterates that colour in natural images is very important for classification. Turning to the performance variation with the number of support regions for dense SIFT. It can be seen that best results are obtained using four concentric circles. With only one support region to represent each patch, results are around 1% worse. This is

Visual Vocabulary	GP	CP	G4CC	C4CC	PS	BOW	GlC	GlT
All categ.	71.51	77.05	84.39	86.65	82.6	82.53	55.12	62.21
Natural categ.	75.43	82.47	88.56	90.28	84.05	88.74	59.53	69.61
Man-made categ.	77.44	83.56	91.17	92.52	89.34	89.67	66.11	73.14

Table 1. Rates obtained different features when using database OT: GP (Grey Patches), CP (Colour Patches), G4CC (Grey SIFT four Concentric Circles), C4CC (Colour SIFT four Concentric Circles), PS (Colour Patches and Colour SIFT), BOW (Bag-of-Words), GlC (Global colour), GlT (Global Texture).

probably because of lack of invariance to scale changes (compared to using four support regions to represent each point).

All the results above are for $P(z|d)$ with the KNN algorithm. Now we investigate classifying the BOW representation directly. We use $V = 1500$, $Z = 25$, $K = 10$, $M = 10$ and four concentric circles. When classifying the 4 natural images in the OT dataset, the results using the topic distribution is 90.28 and with the bag-of-words directly the classification performance decreases by only around $1,5\%$, to 88.74%. However for 8 categories, the performance decreases by nearly 4%, from 86.65 to 82.53%. Using the 13 categories from the FP dataset, the performance falls 8.4%, from 73.4% to 64.8%. Thus there is a clear gain in using pLSA (over the BOW) when classifying a large number of categories.

Table 1 summarizes the results for the three OT image sets (all 8 categories, 4 natural and 4 man-made) covering the different vocabularies: grey and colour patches, grey and colour SIFT, BOW classification and the two baseline algorithms. From these results it can be seen that: (i) The baseline texture algorithm works better than the baseline colour in all three cases. Despite its simplicity the performance of the baseline texture algorithm on man-made images (73.14%) is very high, showing that these images may be easily classified from their edge directions. (ii) 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; colour is always superior to grey level. The best performance (86.65% for all 8 categories) is obtained using colour SIFT with $M = 10$ and four concentric circles. (iii) Somewhat surprizingly, 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. This issue will be investigated further.

The best classified scenes are *highway* and *forest* with 95.61% and 94.86% of correct classified images respectively. The most difficult scenes to classify are *open country*. There is confusion between the *open country* and *coast* scenes. These are also the most confused categories in [11].

Figure 5 shows examples of segmentation of four topics using the colour SIFT vocabulary. Circular patches are painted according to the maximum posterior

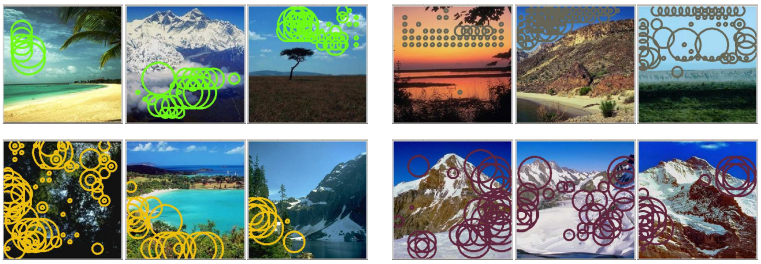


Fig. 5. Topics segmentation. Four topics (clouds – top left, sky – top right, vegetation – lower left, and snow/rocks in mountains – lower right) are shown. Only circular regions with a topic posterior $P(z|w,d)$ greater than 0.8 are shown.

# img. (nt)	2000	1600	1024	512	256	128	32
Perf. $P(z\|d)$	86.9	86.7	84.6	79.5	75.3	68.2	58.7
Perf. BOW	83.1	82.6	80.4	72.8	60.2	52.0	47.3

Table 2. Comparison of $P(z|d)$ and BOW performance as the number of training images used in KNN is decreased. The classification task is into 8 categories from the OT dataset.

$P(z|w,d)$:

$$P(z|w,d) = \frac{P(w|z)P(z|d)}{\sum_{z_l \epsilon Z} P(w|z_l)P(z_l|d)} \tag{2}$$

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 colour, so each colour represents a different topic (the topic colour 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.

Decreasing the number of training images. We evaluate now the classification performance when less training data is available. The OT dataset is split into 2000 training images and 688 test images. A varying number of nt labelled images from the training set are used to learn the pLSA topics and for the KNN. The classification performance is compared using $P(z|d)$ and BOW vectors. The vocabulary has $V = 1500$ words, and $Z = 25$ and $K = 10$. Four support regions are used for each point spaced at $M = 10$. Table 2 shows the results. The gap between pLSA and BOW increases as the number of labelled training images decreases, as was demonstrated in [12].

Summary. The best results are obtained using dense descriptors – colour SIFT with four circular support. Overlap increases the performance. When using the SIFT vocabulary the values for the parameters giving the best results are $M = 10$ pixels with radius for the concentric circles support regions of $r = 4$, 8, 12 and 16 pixels and $V = 1500$, $Z = 25$ and $K = 10$. For patches the best results are

Dataset	# of categ.	our perf.	authors' perf.
OT	8	86.65	–
OT	4 Natural	90.2	89.0 [11]
OT	4 Man-Made	92.5	89.0 [11]
VS	6	85.7	74.1 [19]
FP	13	73.4	65.2 [3]

Table 3. Comparison of our algorithm with other methods using their own databases.

for $N = 5$, $M = 3$, $V = 900$, $Z = 23$ and $K = 10$. In both, colour information is used. The result that dense SIFT gives the best performance was also found by [2] in the case of pedestrian detection. It it interesting that the same feature applies both to more distributed categories (like grass, mountains) as well as the compact objects (pedestrians) of their work where essentially only the boundaries are salient.

6.2 Comparison to previous results

We compare the performance of our classification algorithm to the supervised approaches of Vogel and Schiele [19] and Oliva and Torralba [11], and the semi-supervised approach of Fei Fei and Perona [3], using the same datasets that they tested their approaches on. For each dataset we use the same parameters and type of visual words ($V = 1500$, $Z = 25$ and $K = 10$ with SIFT and four circular supports spaced at $M = 10$). We used colour for OT and VS and grey for FP. The visual vocabulary is computed independently for each dataset, as described in section 5.2. We return to the issue of sharing vocabularies across datasets in section 6.3. The results are given in Table 3.

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. Our method outperforms all of the previous methods, 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 classification step. This is in contrast to [3], where each image is labelled with the identity of the scene to which it belongs during the training stage. In [19], the training requires manual annotation of 9 semantic concepts for 60000 patches, while in [11] training requires manual annotation of 6 properties for thousands of scenes. We are not using the same split into training and testing images as the original authors: for OT we use approximately 200 images per category which means *less* training images (and more testing images) than [11], who used between 250 and 300 training images per category. For VS we used 350 images for training and 350 also for testing which also means *less* training images than [19] who used approximately 600 training images. When working with FP we used 1344 images for training, which is slightly *more* than [3], who used 1300 (100 per category) training images.

Discussion. The superior performance (compared to [3, 19]) could be due to the use of better features and how they are used. In the case of Vogel and Schiele, they learn 9 topics (called *semantic concepts*) that correspond to those that humans can observe in the images: *water*, *trees*, *sky* etc. for 6 categories. Fei Fei and Perona learn 40 topics (called *themes*) for 13 categories. They do not say if these topics correspond to natural objects. In our case, we discover between 22 and 30 topics for 8 categories. These topics can vary depending if we are working with colour features (where topics can distinguish objects with different colours like *light sky*, *blue sky*, *orange sky*, *orange foliage*, *green foliage* etc...) or only grey SIFT features (objects like *trees* and *foliage*, *sea*, *buildings* etc...). In contrast to [19] we discover objects that sometimes would not be distinguished in a manual annotation, for example *water with waves* and *water without waves*. Our superior performance compared to [11] could be due to their method of scene interpretation. They use the spatial envelope modeled in a holistic way in order to obtain the structure (shape) of the scene using coarsely localized information. On the other hand, in our approach specific information about objects is used for scene categorization.

6.3 Other applications

We applied the pLSA based classifier in three other situations. The first one is also a classification task, but combining the images of two different datasets, the second is a relevance feedback application, and the third is scene retrieval for the film *Pretty Woman* [Marshall, 1990]. In all the following the descriptor is dense colour 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).

Vocabulary generalization. In this classification test, we train the system with the four natural scenes of the OT dataset (*coast*, *forest*, *mountains* and *open country*) and test using the same four scene categories from the VS dataset. This tests whether the vocabulary and categories learnt from one dataset generalizes to another. We obtain a performance of 88.27% of correctly classified images. Note, this performance is worse than that obtained when classifying the same categories using only the OT database. This is because (i) images within the same database are more similar, and (ii) the images in VS are more ambiguous and not all represented in OT. To address (i) we will investigate using vocabularies composed from both databases.

Relevance Feedback (RF). [20] proposed a method for improving the retrieval performance, given a probablistic 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 [13]. 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 labelled 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

Fig. 6. Example frames from the film Pretty Woman with their classification. The classifier is trained on the OT dataset.

query images, 25 of each category, in OT dataset. 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* while 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 (8 categories), and key frames from the movie *Pretty Woman* are used as test images. We used one of every 100 frames from the movie to form the testing set. In this movie there are only a few images that could be classified as the same categories used in OT, and there are many images containing only people. So it is a difficult task for the system to correctly classify the key frames. However, the results obtained (see Figure 6) are very encouraging and show again the success of using pLSA in order to classify scenes according to their topic distribution.

7 Conclusions

We have proposed a scene classifier that learns categories and their distributions in unlabelled training images using pLSA, and then uses their distribution in test images as a feature vector in a supervised nearest neighbour scheme. In contrast to previous approaches [3, 11, 19], our topic learning stage is completely unsupervised and we obtain significantly superior performance. 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 objects 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 dataset, together with the optimized features and vocabularies, that is responsible for the superior performance of the scene classifier over previous work (with manual annotation). 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.

Acknowledgements

Thanks to A.Torralba, J.Vogel and F.F.Li for providing their datasets and to Josef Sivic for discussions. This work was partially funded by the research grant BR03/01 from the University of Girona and by the EC NOE Pascal.

References

1. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. SLCV Workshop, ECCV, (2004) 1–22
2. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. CVPR, San Diego, California (2005)
3. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. CVPR, Washington, DC, USA, (2005) 524–531
4. Geodeme, T., Tuytelaars, T., Vanacker, G., Nuttin, M., Van Gool, L. Omnidirectional Sparse Visual Path Following with Occlusion-Robust Feature Tracking. OMNIVIS Workshop, ICCV (2005)
5. Hofmann, T.: Probabilistic latent semantic indexing. ACM SIGIR, (1998)
6. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning **41** (2001) 177–196
7. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using affine-invariant regions. CVPR, volume 2, (2003) 319–324
8. Leung, T., Malik, J.: Representing and recognizing the visual appearance of materials using three-dimensional textons. IJCV **43** (2001) 29–44
9. Lowe, D.: Distinctive image features from scale invariant keypoints. IJCV **60** (2004) 91–110
10. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. IJCV **60** (2004) 63–86
11. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. IJCV (**42**) 145–175
12. Quelhas, P., Monay, F., Odobez, J., Gatica-Perez, D., Tuytelaars, T., Van Gool, L.: Modeling scenes with local descriptors and latent aspects. ICCV, Beijing, China, (2005)
13. Rocchio, J.: Relevance feedback in information retrieval. In the SMART Retrieval System - Experiments in Automatic Document Processing, Prentice Hall, Englewood Cliffs, NJ (1971)
14. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. ICCV, (2003)
15. Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.T.: Discovering objets and their locations in images. In: ICCV, Beijing, China (2005)
16. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. ICCV, Bombay, India (1998) 42–50
17. Vailaya, A., Figueiredo, A., Jain, A., Zhang, H.: Image classification for content-based indexing. T-IP **10** (2001)
18. Varma, M., Zisserman, A.: Texture classification: Are filter banks necessary? CVPR, volume 2, Madison, Wisconsin (2003) 691–698
19. Vogel, J., Schiele, B.: Natural scene retrieval based on a semantic modeling step. CIVR, Dublin, Ireland (2004)
20. Zhang, R., Zhang, Z.: Hidden semantic concept discovery in region based image retrieval. CVPR, Washington, DC, USA (2004)