IDIAP RESEARCH REPORT

Modeling Scenes with Local Descriptors and Latent Aspects

Pedro Quelhas [a] Florent Monay [a]
Jean-Marc Odobez [a] Daniel Gatica-Perez [a]
Tinne Tuytelaars [b] Luc Van Gool [b]

IDIAP–RR 04-79

October 2005

first revision : November 2004
second revision : July 2005

published in
Proc. of IEEE int. Conf. on Computer Vision

[a] IDIAP Research Institute, 1920 Martigny,Switzerland
[b] Katholieke Universiteit, Leuven, Belgium

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-79

Modeling Scenes with Local Descriptors and Latent Aspects

Pedro Quelhas Florent Monay Jean-Marc Odobez Daniel Gatica-Perez
Tinne Tuytelaars Luc Van Gool

October 2005

first revision : November 2004
second revision : July 2005

published in
Proc. of IEEE int. Conf. on Computer Vision

Abstract. We present a new approach to model visual scenes in image collections, based on local invariant features and probabilistic latent space models. Our formulation provides answers to three open questions:(1) whether the invariant local features are suitable for scene (rather than object) classification; (2) whether unsupervised latent space models can be used for feature extraction in the classification task; and (3) whether the latent space formulation can discover visual co-occurrence patterns, motivating novel approaches for image organization and segmentation. Using a 9500-image dataset, our approach is validated on each of these issues. First, we show with extensive experiments on binary and multi-class scene classification tasks, that a bag-of-visterm representation, derived from local invariant descriptors, consistently outperforms state-of-the-art approaches. Second, we show that Probabilistic Latent Semantic Analysis (PLSA) generates a compact scene representation, discriminative for accurate classification, and significantly more robust when less training data are available. Third, we have exploited the ability of PLSA to automatically extract visually meaningful aspects, to propose new algorithms for aspect-based image ranking and context-sensitive image segmentation.

1 Introduction

Scene models are necessary for a number of vision tasks, including classification and segmentation. Among these, scene classification is an important task which helps to provide contextual information to guide other processes such as object recognition [16]. From the application viewpoint, scene classification is relevant in systems for organization of personal and professional imaging collections, and has been widely explored in content-based image retrieval [15, 14, 18, 19]. However, existing approaches are mainly based on global features extracted from the whole image [18, 15] or on fixed spatial layouts [15, 8, 19, 18].

In computer vision, viewpoint invariant local descriptors [10, 6, 17] (i.e. features computed over automatically detected local areas) have proven to be useful in long-standing problems such as viewpoint-independent object recognition, wide baseline matching, and image retrieval. Thanks to their local character, they provide robustness to image clutter, partial visibility, and occlusion. They were designed to have high degree of invariance, and, as a result, are robust to changes in viewpoint and lighting conditions. Recent works have exploited these features to perform retrieval within video [13], or multi-object image categorization [20].

However, scene classification is different than image retrieval [10, 13] or object categorization [20]. While images of a given object are usually characterized by the presence of a limited set of specific visual parts tightly organized into different view-dependent geometrical configurations, a scene is generally composed of several entities (car, house, door, tree, rocks...) organized in often unpredictable layouts. Hence, the content of images from a specific scene type exhibits a large variability. Whereas the specificity of an object might rely on the geometrical configuration of a limited number of visual patterns [13, 5], we expect that the specificity of a particular scene type greatly rests on particular co-occurrences of a large number of visual components.

In this paper, we present an approach to model scenes, and apply it to a number of visual tasks related to scene classification. Our approach integrates the recently proposed scale-invariant feature [6, 10] and probabilistic latent space model [7, 2] frameworks. Our paper describes a number of contributions, both algorithmic and experimental. We first show that invariant local features, represented by bags-of-visterms, are suitable for scene classification, Secondly, we show that PLSA, an unsupervised probabilistic model for collections of discrete data, has the dual ability to generate a robust, low-dimensional scene representation, and to automatically capture meaningful scene aspects. We have successfully used the first property for scene classification, and have exploited the second one to design two new algorithms: one for aspect-based image ranking, and another for context-sensitive image segmentation.

The paper is organized as follows. Section 2 discusses related work. Section 3 presents our approach. Section 4 describes the experiments and results obtained in scene classification. Section 5 reports and discusses the algorithms and results obtained for ranking and segmentation. Section 6 concludes the paper.

2 Related Work

The problem of scene modeling for classification using low-level features has been studied in image and video retrieval for several years [15, 14, 18]. Color, texture, and shape features have been widely used in combination with supervised learning methods to classify images into several semantic classes (indoor, outdoor, city, landscape, sunset, forest, mountain, coastal...). Vogel et al. [19] use both color and texture and a spatial grid layout to perform landscape scene retrieval based on a two-stage retrieval system. The two-stage system makes use of an intermediary semantic level of block classification (concept level) to do retrieval based on the occurrence of such concepts in an image. Graphical models were used by Kumar et al. [8] to detect and localize man made structures in a scene, doing in this way scene segmentation and classification.

The use of local descriptors has become popular for object detection and recognition. Fergus et

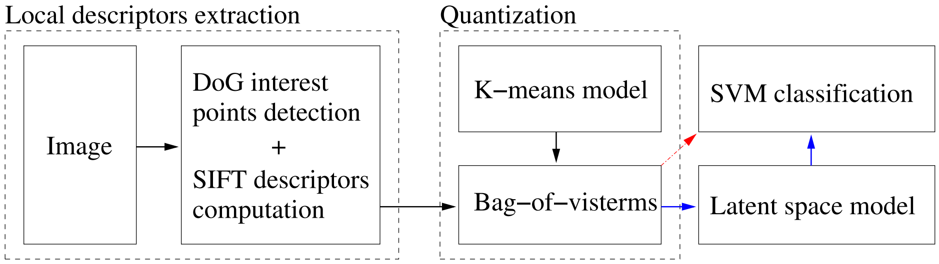

Figure 1: Image representation extraction and classification

al. [5] optimize, in a joint model, a scale-invariant localized appearance model together with a spatial distribution model. Dorko et al. [3] perform feature selection to identify local descriptors relevant to a particular object category, given weakly labeled training images. More recently, Adaboost was proposed to learn classifiers from a set of visual features, including local invariant ones [11]. The analogy between local descriptors and words has been exploited recently [13, 20]. In [13], local invariant features are clustered into 'visterms', which allow to search efficiently through a video for frames of the same object or scene. In [20], good results on object matching and multi-class categorization have been reported, using a system based on a bag-of-words representation built from local invariant features. Finally, variations of latent space models have recently been applied to the problem of modeling annotated images [1], but the methods have relied on other types of features, and have not addressed the classification and segmentation problems as we do here.

In what constitutes the closer work to ours, Fei-Fei and Perona [4] independently proposed two variations of Latent Dirichlet Allocation (LDA) [2] to model scene categories. Their approach also relies on a probabilistic co-occurrence visterm analysis. However, they do not study the effect of the amount of less training data and also do not apply their method to obtain segmentation. Furthermore, in their method the introduced class node does not allow model learning using unlabeled data.

3 Image Representation

In this section, we present the two models that will be used as image representation: first the bag-of-visterms (BOV), built from automatically extracted and quantized local descriptors. The second is obtained through higher-level abstraction of the bag-of-visterms into a set of aspects using the latent space modeling.

3.1 Bag-of-visterms Representation

The construction of the BOV feature vector h from an image d involves the different steps illustrated in Fig. 1. In brief, regions of interest are automatically detected in the image, then local descriptors are computed over those regions. All the descriptors are quantized into visterms, and the occurrences in the image of each specific visterm in the vocabulary are counted to build the BOV. In the following we describe in more detail each of the steps.

The goal of the interest point detector is to automatically extract characteristic points in the image which are invariant to some geometric and photometric transformations. From existing detectors [6, 10, 17], we used the difference of Gaussians (DOG) point detector [6]. This detector essentially identifies blob-like regions and is invariant to translation, scale, rotation, and constant illumination variations. We preferred this detector over fully affine-invariant ones [10, 17], as the increase of the invariance degree may remove valuable information about local image content.

Local descriptors are computed on the characteristic region around each detected interest point. We use the SIFT (Scale Invariant Feature Transform) descriptor [6]. This orientation invariant descriptor is based on the grayscale representation of images, and was shown to perform best in terms

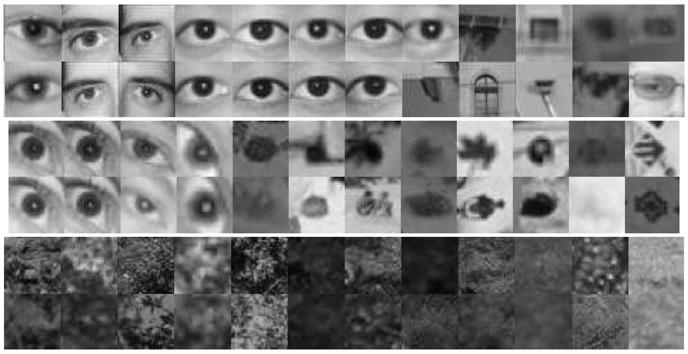

Figure 2: Samples from 3 different visterms. More results in www.idiap.ch/∼monay/ICCV05/.

of specificity of region representation and robustness to image transformations [9]. SIFT features are local histograms of edge directions computed over different parts of the interest region. In [6], it was shown that the use of 8 orientation directions and a grid of 4x4 parts give best results, leading to a descriptor s of size 128.

In order to obtain a text-like representation, we quantize each local descriptor s into one of a discrete set $\mathcal{V}$ of visterms v according to a nearest neighbor rule:

$$s \mapsto Q(s) = v_i \leftrightarrow \text{dist}(s, v_i) \leq \text{dist}(s, v_j), \forall j \in \{1, .., N_{\mathcal{V}}\},$$

where $N_{\mathcal{V}}$ denotes the size of the visterm set. We will call **vocabulary** the set $\mathcal{V}$ of all the visterms. The vocabulary construction is performed through clustering. More specifically, we apply the k-means algorithm to a set of local descriptors extracted from training images, and keep the means as visterms. We used the Euclidean distance in the clustering and quantization processes, and choose the number of clusters depending on the desired vocabulary size.

Finally, the bag-of-visterms (BOV) representation is constructed from local descriptors according to:

$$h(d) = (h_i(d))_{i=1..N_{\mathcal{V}}}, \text{ with } h_i(d) = \text{n}(d, v_i), \tag{1}$$

where $\text{n}(d, v_i)$ denotes the number of occurences of visterm v_i in image d. This vector-space representation of an image contains no information about spatial relationships between visterms, in the same way the standard bag-of-words text representation removes the word ordering information.

3.2 Latent Space Representation

The bag-of-visterms representation is simple to build. However it may suffer from two issues: *polysemy* -a single visterm may represent different scene content- and *synonymy* -several visterms may characterize the same image content. To illustrate these issues, consider samples from three different visterms obtained when building the vocabulary V_{1000} (see Section 4.4 for details), as shown in Figure 2. As can be seen, the top visterm (first two rows) represents mostly eyes. However, windows and publicity patches get also indexed by this visterm, indicating the polysemic nature of that visterm, which means that although this visterm will mostly occur on faces, it can also occur in city environments. The second two rows present samples from another visterm. Clearly, this visterm also represents eyes, which makes it a synonym of the first displayed visterm. Finally, the samples of a third visterm (last two rows) indicate that this visterm captures a certain fine grain texture arising from different contexts, illustrating that not all visterms have a clear semantic interpretation.

Recently, probabilistic latent space models [7, 2] have been proposed to capture co-occurrence information between elements in a collection of discrete data in order to disambiguate the bag-of-words representation. The analysis of visterm co-occurrence can be considered using similar approaches. In this paper, we use the Probabilistic Latent Semantic Analysis [7] model for that purpose.

PLSA is a statistical model that associates a latent variable $z_l \in \mathcal{Z} = \{z_1, \ldots, z_{N_A}\}$ with each observation (the occurrence of a word in a document). These variables, usually called aspects, are then used to build a joint probability model over images and visterms, defined as the mixture

$$P(v_j, d_i) = P(d_i) \sum_{l=1}^{N_A} P(z_l|d_i) P(v_j|z_l). \tag{2}$$

PLSA introduces a conditional independence assumption: it assumes the occurrence of a visterm v_j to be independent of the image d_i it belongs to, given an aspect z_l. The model in Equation 2 is defined by the conditional probabilities $P(v_j|z_l)$ which represent the probability of observing the visterm v_j given the aspect z_l, and by the image-specific conditional multinomial probabilities $P(z_l|d_i)$. The model expresses the conditional probabilities $P(v_j|d_i)$ as a convex combination of the aspect specific distributions $P(v_j|z_l)$.

The parameters of the model are estimated using the maximum likelihood principle, using a set of training images $\mathcal{D}$. The optimization is conducted using the Expectation-Maximization (EM) algorithm [7]. This estimation procedure allows to learn the aspect distributions $P(v_j|z_l)$. These image independent parameters can then be used to infer the aspect mixture parameters $P(z_l|d)$ of any image d given its BOV representation $h(d)$. Consequently, the second image representation we will use is defined by:

$$a(d) = (P(z_l|d))_{l=1\ldots N_A} \tag{3}$$

This representation will be used as input to a scene classifier as well as to perform the visterm image segmentation.

4 Scene Classification Results

Experiments are divided into three separate problems: indoor/outdoor, city/landscape, and indoor/city/landscape. In this section we describe the datasets used, protocol and baseline setup. We show classification results of our approach, first using the BOV representation, then using the aspect representation, and compare them with the baseline method. We also analyze the evolution of the results under different conditions (vocabulary size, number of latent aspects, amount of training data).

4.1 Datasets and Protocol

In our experiments, we used 4 datasets.

D1: a subset of the COREL [18] database composed of 2505 city and 4175 landscape images, of size 384x256 pixels.

D2: composed of 2777 indoor images retrieved from the Internet. The size of these images are approximately 384x256 pixels. Images with larger size were resized using bilinear interpolation. Image size in the dataset was kept similar since, it is known that the number of detected interest points is highly dependent on the resolution of the image and would bias the bag-of-visterms representation.

D3: constituted by 3805 images from several sources: 1002 building images (ZuBud) [12], 144 images of people and outdoors [11], 435 indoor peoples' faces [20], 490 indoors (COREL) [18], 1516 city/landscape overlap images (COREL) [18] and 267 Internet photographic images.

D4: composed of all images taken from the datasets **D1** and **D2**.

We use the dataset **D1** for city/landscape scene classification, and **D4** for indoor/outdoor and indoor/city/landscape scene classification. Dataset **D3** was used for vocabulary construction. Using 3805 images, we obtained approx. 1 million descriptors for vocabulary construction.

Method	indoor/outdoor		city/landscape	
baseline	10.4	(0.8)	8.3	(1.5)
BOV V_{100}	8.5	(1.0)	5.5	(0.8)
BOV V_{300}	7.4	(0.8)	5.2	(1.1)
BOV V_{600}	7.6	(0.9)	5.0	(0.8)
BOV V_{1000}	7.6	(1.0)	5.3	(1.1)

Table 1: Classification error for the baseline model and the BOV representation, for 4 different vocabularies. Mean and standard deviation (in brackets) are shown.

The protocol of all classification experiments was as follows. The full dataset was divided into 10 parts, resulting in 10 different splits of the full dataset. One split corresponds to keeping one part of the data for testing, while using the other 9 parts for training. In this way, we obtain 10 different classification results. Reported values for all experiments correspond to the average error over all splits, and standard deviations of the errors are provided in brackets after the mean value.

These datasets are also used for further experiments with latent aspect models in the next section.

4.2 Baseline Methods

We use Vailaya et al. [18] methods as baseline. We chose these methods since they allow for good classification on landscape, city and indoor scenes, and are commonly regarded as the most representative state-of-the-art baseline. A different strategy is used to tackle each problem. Color features are used to classify images as indoor/outdoor, and edge features are used to classify outdoor images as city/landscape. Color features are based on the collection of the LUV first and second order moments in a 10x10 spatial grid in the image, resulting in a 600-dimensional histogram feature. Edge features are based on edge coherence histograms calculated on the whole image. Edge coherence histograms are based on extracting edges in coherent neighborhoods, eliminating areas where edges are noisy. Directions are then discretized into 72 directions and placed on a histogram. An extra non-edge pixels bin is added to the histogram. The final feature's dimension is 73.

The baseline approach applies both baseline methods in a hierarchical implementation. Images are classified as indoor or outdoor based on color, and next all correctly classified outdoor images are classified as city or landscape based on an edge coherence direction histogram.

4.3 SVM Classifier

To classify an input image d represented either by the bag-of-visterms vector h, aspect parameters a, or any of the baseline's feature vector (see previous section), we employed Support Vector Machines (SVMs). We used Gaussian kernel SVMs. Hyperparameters of the SVM (e.g. the bandwidth) were chosen based on a 5-fold cross-validation.

4.4 Results and Discussion BOV

To analyze the effect of varying the vocabulary size employed to construct the BOV representation, we considered four vocabularies of 100, 300, 600 and 1000 visterms, denoted by V_{100}, V_{300}, V_{600}, and V_{1000}, respectively, constructed from **D3** as described in Section 3.

Binary Classification

Table 1 provides the classification error for the binary classification tasks. First, we can see that the BOV approach consistently outperforms the baseline methods. This is confirmed by the Paired T-test criterion in all cases, for p=0.05.

Regarding vocabulary size we can see that for vocabularies of 300 visterms or more the classification errors are equivalent. This contrasts with the work in [20], where the 'flattening' of the classification

Method	indoor/city/landscape
baseline	15.9 (1.0)
BOV V_{100}	12.3 (0.9)
BOV V_{300}	11.6 (1.0)
BOV V_{600}	11.5 (0.9)
BOV V_{1000}	11.1 (0.8)
BOV V_{1000} hier.	11.1 (1.1)

Table 2: Three class classification error for baseline and BOV models. The baseline model system is hierarchical (cf Section 5.2).

Total Classification error				11.1 (0.8)
	Resulting Classification			Classification
Ground Truth	indoor	city	landscape	Error (%)
indoor	2489	242	23	10.3
city	364	1873	268	25.2
landscape	49	84	4042	3.1

Table 3: Confusion matrix from the three-class classification problem, using vocabulary V_{1000}. The total number of classified images is presented.

performance was observed only from a vocabulary of 1000 visterms. A possible explanation may come from the difference in task (they perform object image classification) and in their use of the Harris-Affine point detector [10]. The DOG point detector is known to be more stable than the Harris-Affine detector [9].

The results of Table 1 show that constructing a vocabulary from an auxiliary dataset **D3** does well in our experiments. This suggests that as long as the auxiliary dataset contains significant images for our task, it allows to build a good visterm vocabulary. This point is especially relevant in practice, as it could allow for re-usability if we find a dataset that is significant for several tasks.

Three-class Classification

Combining both classification problems, we define a 3-class classification problem (indoor/city/landscape). We present results with BOV in Table 2 along with the baseline. Classification results were obtained using both a multi-class SVM and two binary SVMs in the hierarchical case.

First, we can see that once again our system outperforms the state-of-the-art approach with statistically significant differences. Secondly, we again observe the stability of results with vocabularies with 300 or more visterms, the vocabulary of 1000 visterms giving slightly better performance. Based on these results, we assume V_{1000} to be optimal and use it for all remaining experiments in this paper.

For the 3-class classification experiments we can further analyze results by looking at the confusion matrix, in Table 3. We see that landscape images are well classified, and indoor images get slightly confused with city images. However, performance lowers for city images, which get classified as both indoor and landscape. This may be caused by the fact that city images often contain visterms that can also occur in images of other classes.

4.5 Results and Discussion PLSA

In PLSA, we use the probability of each latent aspect l given each specific document i $P(z_l|d_i)$ as a N_A dimensional feature vector. Without any reference to the class label during the PLSA model learning, how much discriminant information would remain in this aspect representation? To evaluate

PLSA-I	A	indoor/outdoor	city/landscape	3-class
V_{1000}	20	9.5 (1.0)	5.5 (0.9)	12.6 (0.8)
V_{1000}	60	8.3 (0.8)	4.7 (0.9)	11.2 (1.3)
PLSA-O	A	indoor/outdoor	city/landscape	3-class
V_{1000}	20	8.9 (1.4)	5.6 (0.9)	12.3 (1.2)
V_{1000}	60	7.8 (1.2)	4.7 (0.9)	11.8 (1.0)

Table 4: Comparison of PLSA-I and PLSA-O strategies on the indoor/outdoor and city/landscape scene classification tasks, using 20 or 60 aspects and for vocabulary V_{1000}.

PLSA-O					
A	20	40	60	80	100
Error	5.6 (0.9)	4.9 (0.8)	4.7 (0.9)	4.8 (1.0)	5.0 (0.9)

Table 5: Classification results for city/landscape using different number of aspects for PLSA-O.

this, we compare the classification errors obtained with the PLSA and BOV representations.

Furthermore we test the influence of the training data on the aspect model. To investigate the latter issue, we conducted two experiments which only differ in the data used to train the aspect models (i.e. the $P(v_j|z_l)$ multinomial probabilities).

PLSA-I : for each split of the data, the training data split (that is used to train the SVM classifier, cf Section 4) was also used to learn the aspect models.

PLSA-O : the aspect models are trained only once on the auxiliary database **D3**.

As the dataset **D3** comprises city, outdoor, and city-landscape overlap images, PLSA performed on this set should capture valid latent aspects for the three classification tasks simultaneously. Such a scheme presents the advantage of constructing a common N_A-dimensional representation for each image that can be tested on all classification tasks.

Classification Results

We show in Table 4 results for PLSA with 20 and 60 aspects, for the PLSA-I and PLSA-O strategies, using V_{1000}. Overall, the performance of PLSA-I and PLSA-O is comparable for city/landscape scene classification, and PLSA-O even significantly improves over PLSA-I for indoor/outdoor. This suggest that learning the aspect model on the same set used for the classifier training may cause some overfitting. Using PLSA we obtain a dimensionality reduction with a factor of 50 and 17 times for 20 and 60 aspects respectively, while keeping the discriminant information and still performing significantly better than the baseline. Since using PLSA-O allows us to learn one single model for several tasks we keep this model for the rest of the paper.

Table 5 displays the evolution of the error with the number of aspects. Results show that the performance is relatively independent of the number of aspects for the city/landscape case. For the rest of this paper we will use a PLSA model with 60 aspects.

Decreasing the Amount of Training Data

Since PLSA captures co-occurrence information from the data from which it is learned, it can provide a more stable image representation. We expect this to help in the case of lack of sufficient training data. Table 6 compares classification errors for the BOV and the PLSA representations for the different tasks when using less data to train the SVMs.

Table 6 shows that PLSA performs better than both baseline and BOV approaches for all reduced training set experiments and deteriorates less as the training set is reduced. Previous work on latent space modeling has reported similar behavior for text data [2]. PLSA better performance in this case is due to the ability of PLSA to capture aspects that contain general information about visual

Data size	90%	10%	5%	2.5%
Indoor/Outdoor				
data size	8511	945	472	236
PLSA	7.8(1.2)	9.1(1.3)	10.0(1.2)	11.4(1.1)
BOV	7.6(1.0)	9.7(1.4)	10.4(0.9)	12.2(1.0)
Baseline	10.4(0.8)	15.9(0.4)	19.0(1.4)	23.0(1.9)
City/Landscape				
data size	6012	668	334	167
PLSA	4.7(0.9)	5.8(0.9)	6.6(0.8)	8.1(0.9)
BOV	5.3(1.1)	7.4(0.9)	8.6(1.0)	12.4(0.9)
Baseline	8.3(1.5)	9.5(0.8)	10.0(1.1)	11.5(0.9)
Indoor/City/Landscape				
data size	8511	945	472	236
PLSA	11.8(1.0)	14.6(1.1)	15.1(1.4)	16.7(1.8)
BOV	11.1(0.8)	15.4(1.1)	16.6(1.3)	20.7(1.3)
Baseline	15.9(1.0)	19.7(1.4)	24.1(1.4)	29.0(1.6)

Table 6: Comparison between BOV results and PLSA-O approach, with SVM classifier trained with progressively less training data.

co-occurrence. Thus, while the lack of data impairs the simple BOV representation in covering the manifold of documents belonging to a specific scene class, PLSA-based representation is less affected.

Since we learn the aspect-based representation on auxiliary non-labeled data, the improvement we obtained for reduced training data demonstrates the potential of this approach in partially labeled data problems.

5 PLSA-based Ranking/Segmentation

As shown above, PLSA modeling can improve the classification performance under limited labeled data conditions. However, latent space models were introduced to solve ambiguity issues (cf Section 3) in text modeling, and it is known that the latent structure identified by PLSA relates to the semantic aspects of the data [7]. In this section, we illustrate this relationship on our visual data through two applications: unsupervised image ranking and context based image segmentation.

5.1 Aspect-based Image Ranking

Given an aspect z, images can be ranked according to:

$$P(d|z) = \frac{P(z|d)P(d)}{P(z)} \propto P(z \mid d) \tag{4}$$

The observation of the top-ranked images of an aspect illustrates its potential 'semantic meaning' for a given set of images. Figure 3 displays the 7 most probable images from the first split of the **D1** database, for five out of 20 aspects learned on **D3**. The top-ranked images of aspect 1 and 6[1] belong to the *landscape* class. More precisely, aspect 1 seems to be related to horizon/panoramic scenes, and aspect 4 to forest/vegetation. Conversely, top-ranked images from aspect 4 and 14 are related to the *city* class. However, as aspects are identified by analyzing the co-occurrence of local visual patterns, aspect may be consistent from this point of view (e.g. aspect 19 is consistent in terms of texture) without allowing for a direct semantic interpretation.

[1] Note that the aspect indices have no intrinsic relevance to a given class, given the unsupervised nature of the PLSA model learning.

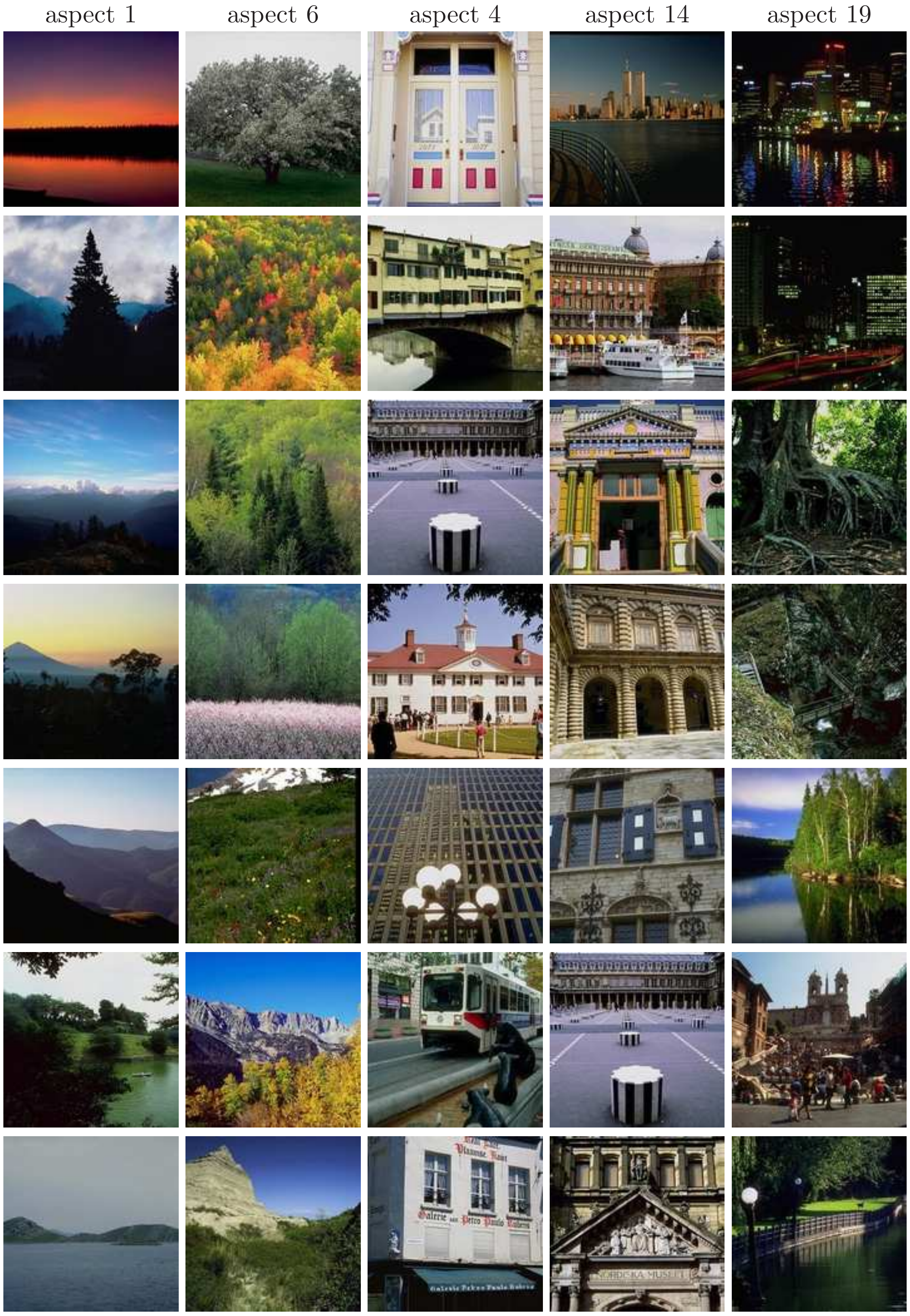

Figure 3: The 7 most probable images from dataset **D1** for five aspects out of 20 learned on **D3** images. More results in www.idiap.ch/∼monay/ICCV05/.

The correspondence between aspects and scene type can be measured more objectively by considering the aspect-based image ranking as an image retrieval system. Defining the *Precision* and *Recall* paired values by:

$$Precision(r) = \frac{RelRet}{Ret} \quad Recall(r) = \frac{RelRet}{Rel},$$

where Ret is the number of retrieved images, Rel is the total number of relevant images and $RelRet$ is the number of retrieved images that are relevant, we can compute precision/recall curves associated with each aspect-based image ranking considering either *City* and *Landscape* queries, as illustrated in Fig. 4. Those curves demonstrate that some aspects are related to either 'City' or 'Landscape' concept, and confirm observations made previously with respect to aspects 4, 6 and 14. As expected, aspect 19 does not appear in either the *City* or *Landscape* top precision/recall curves. These results illustrate that the latent data structure identified by PLSA correlates with the semantic structure of our data and makes PLSA a potential tool for browsing/annotating unlabelled image dataset.

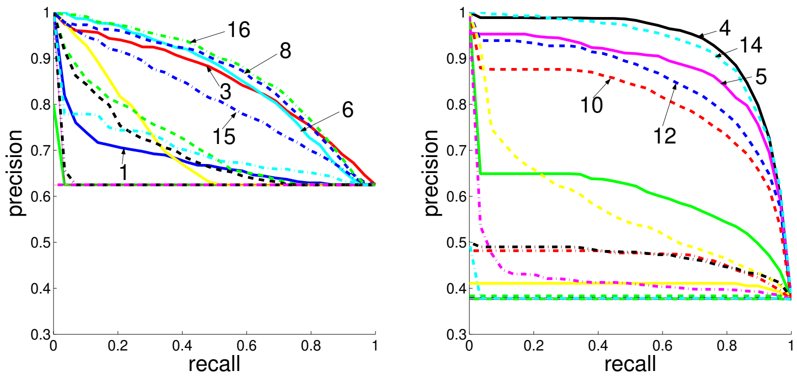

Figure 4: Precision/recall curves for each of the 20 aspect-based image rankings, relative to the landscape (left) and city (right) query. Floor precision values correspond to the proportion of city (resp. landscape) images in the dataset.

5.2 Aspect-based Image Segmentation

A third way to assess the relevance of the PLSA modeling is to evaluate whether the aspect's individual visterms themselves match the aspect scene type. This can be achieved by mapping each visterm of an image to its most probable aspect and displaying the resulting visterm labeling to generate a segmentation-like image. Accordingly, the mapping can be computed by:

$$\begin{aligned} z_{v_j} &= \arg\max_z (P(z \mid v_j, d_i)) \\ &= \arg\max_z (\frac{P(v_j \mid z)P(z \mid d_i)}{\sum_z P(v_j \mid z)P(z \mid d_i)}). \end{aligned} \tag{5}$$

Fig.5 shows two images along with their aspect distribution over the 20 aspects learned on **D3**. Based on this distribution, the most probable aspects are selected, and only visterms labeled with those aspects are displayed. In Fig. 5(b), aspects 6 and 12 are the most probable, which are related to landscape and city respectively. In the second example, in addition to city and landscape aspects, visterms associated with aspect 1 clearly corroborate its horizon/panoramic semantic meaning suggested by Fig.3. These results show that PLSA modeling not only correctly describes images as mixtures of city- and lanscape-related concepts, but also that visterms labeled by those aspects are located on corresponding image regions.

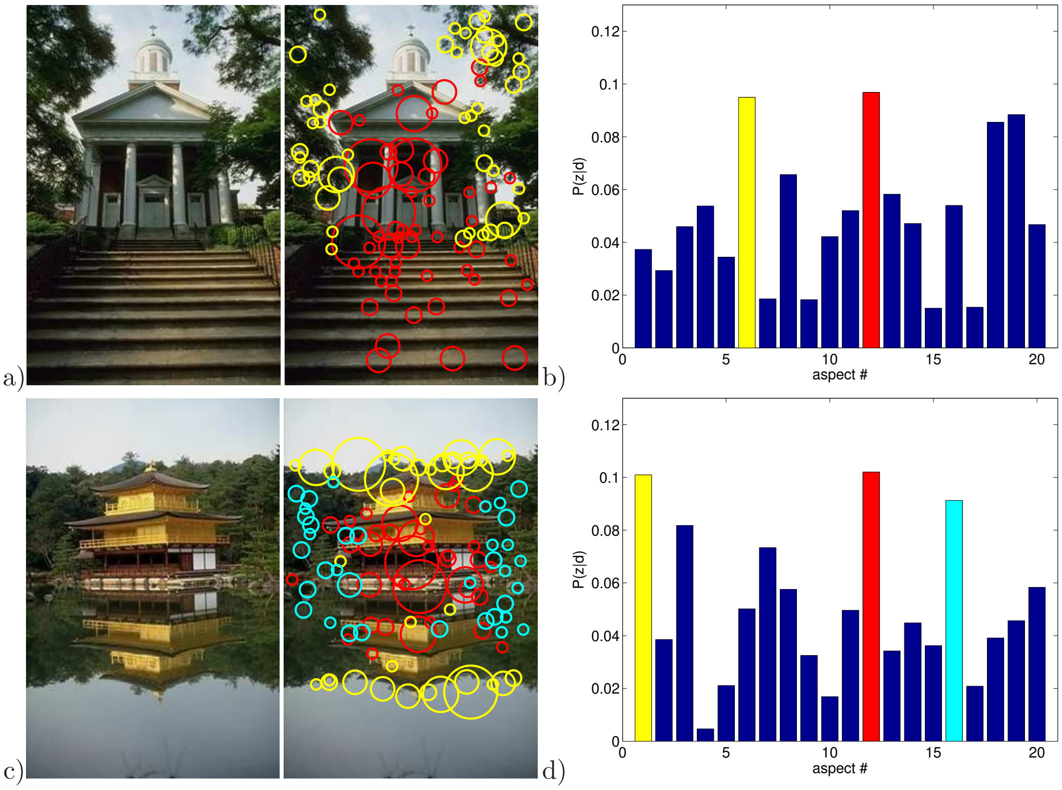

Figure 5: Visterm labeling according to $\arg\max_z(P(z|v,d))$ (see text for details). More results in www.idiap.ch/∼monay/ICCV05/.

A first evaluation of PLSA's segmentation potential is proposed here as a preliminary study. Note however that since visterms are not covering the whole image uniformly by construction, since we attribute visterms to a given region type, the result is a 'sparse segmentation' of the image. For evaluation, we manually segmented 485 images containing both landscape and man-made structures. Then, visterms were attributed to the man-made class if their center fell into the man-made image region, or to the lanscape class otherwise. As evaluation procedure, we considered again precision and recall measures based on 'visterm retrieval', where the task is to retrieve local descriptors related either to man-made structures or landscape.

Given a labeling of all visterms into aspects, different retrieval points are obtained by introducing one aspect at a time and adding its associated visterms in the retrieved list. The introduction order for the man-made (resp. landscape) visterm retrieval task is selected by ranking the aspects according to the average precision of the 'city' (resp. landscape) precision/recall curves in Fig. 4, enabling to successively select the corresponding local descriptors according to the confidence they belong to man-made structures.

We compare two strategies for mapping aspects to local descriptors. The first one, given by Eq. 5, is image-contextual in the sense that the mapping actually depends on the content of image d_i. The second is non-contextual, and consists of building an image-independent mapping by attributing aspects to local descriptors according to:

$$z_{v_j} = \underset{z}{\operatorname{argmax}}\ (P(z|v_j)) = \underset{z}{\operatorname{argmax}}\ (\frac{P(v_j|z)P(z)}{P(v_j)}) \tag{6}$$

By comparing the two mapping methods, we can analyze the effect of learning $P(z|d)$ for a given image and observe if it improves the local descriptors attribution. As can be seen from Figure 6,

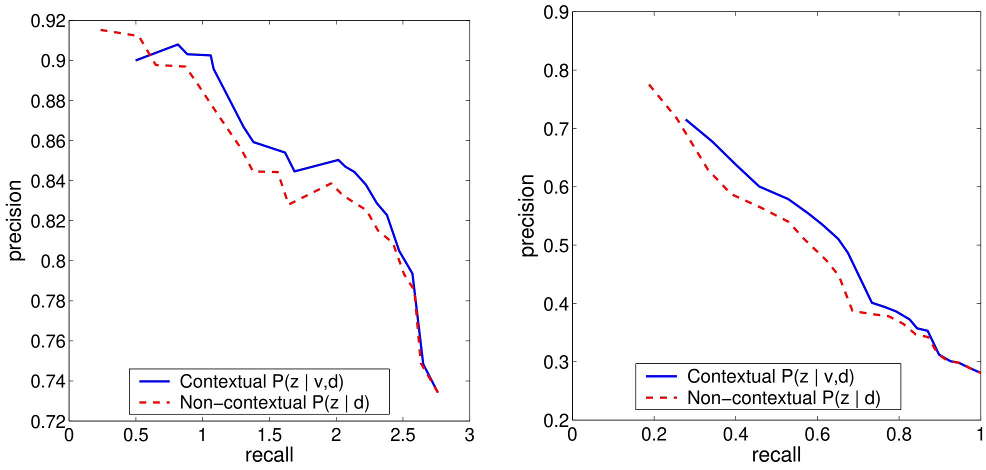

Figure 6: Visterm retrieval precision/recall curves, relative to the landscape (left) and man-made (right) queries, using contextual or non-contextual mapping.

which displays precision and recall curves corresponding to both methods, the introduction of context in the aspect-based segmentation significantly improves the segmentation precision for a given recall.

6 Conclusion

Based on the results presented in this paper, we believe that our scene modeling methodology is promising. We have first shown that the bag-of-local-descriptor approach is adequate for scene classification, consistently outperforming state-of-the-art methods relying on a suite of hand-picked features. We have also shown that the PLSA-based representation is competitive with the bag-of-visterms in terms of performance, but it also provides a number of interesting advantages, including a more graceful performance degradation with decreasing amount of training data, and the multi-faceted clustering property that we have exploited for aspect-based image ranking and contextual image segmentation. Each of these results have value on their own.

One can argue whether the discrete representation obtained with k-means clustering is actually a true 'visual vocabulary'. Visual inspection of the clusters shows they contain meaningful features (e.g., the eyes shown in Fig. 2), but also, in most cases, a lot of noisy patches. This is due to the fact that k-means actually partitions the data, assigning each and every feature to the closest cluster, even if this cluster is relatively far. We plan to study other clustering algorithms, that are equally well suited for high-dimensional data and large datasets, but yield more meaningful clusters. Other paths for further investigation include ways to determine the optimal vocabulary size and/or the number of aspects.

The description of visual scenes as a mixture of aspects is a concept worth of further exploration. We plan to extend our work on scene segmentation based on this concept. We will also study feature fusion mechanisms (e.g. color and local descriptors) in the latent space framework.

Acknowledgments

This work was funded by the PASCAL pump-priming project CARTER. The first four authors were also supported by the Swiss NCCR IM2. T. Tuytelaars was also supported by the Fund for Scientific Research Flanders.

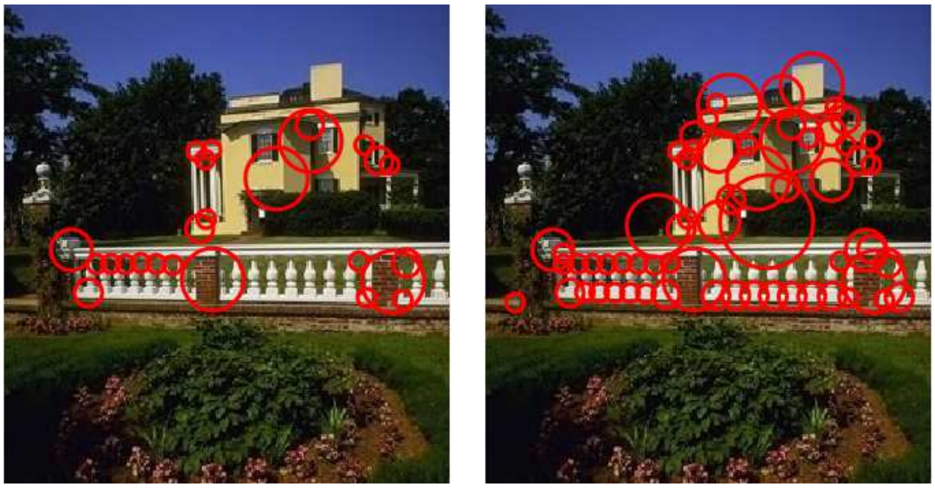

Figure 7: Retrieved visterms of the most relevant city aspect using the two mapping strategies: non-contextual (left) and contextual (right). More results in www.idiap.ch/∼monay/ICCV05/

References

[1] K. Barnard, P. Duygulu, N. Freitas, D. Forsyth, D. Blei, and M.I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[2] D. Blei, Y. Andrew, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1020, 2003.

[3] G. Dorko and C. Schmid. Selection of scale invariant parts for object class recognition. In *Proc. ICCV*, page 634, 2003.

[4] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR*, 2005.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE CVPR*, pages 264–271, 2003.

[6] D. G.Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2003.

[7] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learning*, 42:177, 2001.

[8] Sanjiv Kumar and Martial Herbert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Proc. ICCV*, 2003.

[9] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proc. IEEE CVPR*, page 257, June 2003.

[10] K. Mikolajczyk and C. Schmid. Scale and affine interest point detectors. *IJCV*, 60(1):63–86, 2004.

[11] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proc. ECCV*, pages 41–44, 2004.

[12] H. Shao, T. Svoboda, V. Ferrari, T. Tuytelaars, and L. Van Gool. Fast indexing for image retrieval based on local appearance with re-ranking. In *Proc. IEEE ICIP*, 2003.

[13] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, pages 1470–1477, 2003.

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

[15] M. Szummer and R.W. Picard. Indoor-outdoor image classification. In *IEEE Int. Workshop CAIVD, in ICCV'98*, pages 42–51, 1998.

[16] A.B. Torralba, K.P. Murphy, W.T. Freeman, and M.A. Rubin. Context-based vision system for place and object recognition. In *Proc. ICCV*, 2003.

[17] T. Tuytelaars and L. Van Gool. Content-based image retrieval based on local affinely invariant regions. In *Proc. Visual99*, pages 493–500, 1999.

[18] A. Vailaya, M. Figueiredo, A. Jain, and H.J. Zhang. Image classification for content-based indexing. *IEEE Trans. on Image Processing*, 10(1):117–130, Jan. 2001.

[19] Julia Vogel and Bernt Schiele. Natural scene retrieval based on a semantic modeling step. In *Proc. CIVR*, 2004.

[20] J. Willamowski, D. Arregui, G. Csurka, C.R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *Proc. LAVS04*, August 2004.