A Thousand Words in a Scene

Pedro Quelhas, Florent Monay, Jean-Marc Odobez, Daniel Gatica-Perez and Tinne Tuytelaars

***Abstract*— This paper presents a novel approach for visual scene modeling and classification, investigating the combined use of text modeling methods and local invariant features. Our work attempts to elucidate (1) whether a text-like *bag-of-visterms* representation (histogram of quantized local visual features) is suitable for scene (rather than object) classification, (2) whether some analogies between discrete scene representations and text documents exist, and (3) whether unsupervised, latent space models can be used both as feature extractors for the classification task and to discover patterns of visual co-occurrence. Using several data sets, we validate our approach, presenting and discussing experiments on each of these issues. We first show, with extensive experiments on binary and multi-class scene classification tasks using a 9500-image data set, that the *bag-of-visterms* representation consistently outperforms classical scene classification approaches. In other data sets we show that our approach competes with or outperforms other recent, more complex, methods. We also show that Probabilistic Latent Semantic Analysis (PLSA) generates a compact scene representation, discriminative for accurate classification, and more robust than the *bag-of-visterms* representation when less labeled training data is available. Finally, through aspect-based image ranking experiments, we show the ability of PLSA to automatically extract visually meaningful scene patterns, making such representation useful for browsing image collections.**

***Index Terms*— Image representation, scene classification, object recognition, quantized local descriptors, latent aspect modeling.**

I. Introduction

Scene classification is an important task in computer vision. It is a difficult problem, interesting in its own right, but also as a means to provide contextual information to guide other processes such as object recognition [39]. From the application viewpoint, scene classification is relevant in systems for organization of personal and professional image and video collections. As such, this problem has been widely explored in the context of content-based image retrieval [38], [37], [41], but existing approaches have traditionally been based on global features extracted on the whole image, on fixed spatial layouts, or on image segmentation methods whose results are often difficult to predict and control [5], [38], [41], [31], [15], [16], [42].

In a different direction, viewpoint invariant local descriptors (i.e. features computed over automatically detected local areas) have proven to be useful in long-standing problems such as viewpoint-independent object recognition [7], [44], [27], wide baseline matching [21], [40], [19] and, more recently, in image retrieval [34], [12]. Thanks to their local character, they provide robustness to image clutter, partial visibility, and occlusion. Thanks to their invariant nature, changes in viewpoint can be dealt with in a natural way, while providing robustness to changes in lighting conditions. All these properties make the features stable, producing a relatively repeatable representation of a particular object. In the case of scenes, since we expect the component parts of a given scene class to have relatively similar image representations, these features could potentially be useful to detect and describe similar local scene areas consistently, thus providing good generalization properties.

In a sense, these local invariant features show many commonalities with the role played by words in traditional document analysis techniques [1], in that they are local, have a high repeatability between similar images of similar scenes, and have a relatively high discriminant power. This analogy has been exploited in recent works to perform retrieval within videos [34], or object classification [44], and is studied here in more detail.

However, scene classification is clearly different from image retrieval and object categorization. On one hand, 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. On the other hand, a scene is generally composed of several entities (e.g. car, house, building, face, wall, door, tree, forest, rocks), organized in often unpredictable layouts. Hence, the visual content (entities, layout) of a specific scene class exhibits a large variability, characterized by the presence of a large number of different visual descriptors. In view of this, while the specificity of an object strongly relies on the geometrical configuration of a relatively limited number of visual descriptors [34], [12], the specificity of a scene class greatly rests on the particular patterns of co-occurrence of a large number of visual descriptors.

In this paper, we propose a novel approach for scene classification that integrates scale-invariant feature extraction and latent space modeling methods. The contributions of our paper are the following.

1) An approach for scene classification, based on the use of *bags-of-visterms* (BOV) (i.e. quantized invariant local descriptors) to represent scenes. Even though recent work used quantized local descriptors for object matching in videos [34], and for object classification [44], our work demonstrates that this approach is successful to classify scenes. We show this by presenting extensive experiments on two binary and four multi-class classification tasks (including 3, 5, 6, and 13 classes). Moreover, we show by a rigorous comparison that our work consistently outperforms classical scene classification approaches [41]. We also show that our approach is clearly competitive when compared to approaches that have recently appeared [42] or that have been developed in parallel to ours [11]. Finally, to provide new insights about the analogy between the bag-of-visterms representation and text, we have conducted a study of sparsity, co-occurrence, and discriminative power of visterms, which complements and extends the work by [34], in a different media source.

2) A novel approach for scene classification, based on the use of probabilistic latent space models [14], [3] that have proven to be successful in text modeling, to build scene representations beyond the bag-of-visterms. Latent space models capture co-occurrence information between elements in a collection of discrete data that simpler representations usually cannot, and allow to address issues related to synonymy (different visterms may represent the same scene type) and polysemy (the same visterm may represent different scene types in different contexts), which can be encountered in scene classification. We show that Probabilistic Latent Semantic Analysis (PLSA) allows for the extraction of a compact, discriminant representation for accurate scene classification, that outperforms global scene representations, and remains competitive with recently proposed approaches. This compact representation is especially robust when labeled training data is scarce, and allows for a greater re-usability of our framework, as labeling is a time-consuming task. All of our findings are based on extensive experiments. Although related, the approach we propose differs from the ones discussed in [11] for scene classification and [33] for object clustering. A detailed discussion of the differences is presented in the next Section.

3) A novel approach for scene ranking and clustering, based on the successful use of the PLSA formulation. We show that PLSA is able to automatically capture meaningful scene aspects from data, where scene similarity is evident, which makes our PLSA-derived representation useful to explore the scene structure of an image collection, and thus turning it into a tool with potential in visualization, organization, browsing, and annotation of images in large collections.

The rest of the paper is organized as follows. The next Section discusses related work. Section III presents the image representations we explore. Section IV compares properties of these representations with text document representations. Section V describes the classifier we use. Section VI presents our experimental setup. Classification results are provided and discussed in Section VII. Section VIII describes the aspect-based image ranking results. Section IX compares our method with recently proposed works, on other existing scene classification data sets. Section X concludes the paper.

II. Related Work

The problem of scene classification using low-level features has been studied in image and video retrieval for several years [13], [38], [41], [26], [25], [28], [37]. Broadly speaking, the existing methods differ by the definition of the target scene classes, the specific image representations, and the classification method. We focus the discussion on the first two points. With respect to scene definition, most methods have aimed at classifying images into a small number of semantic scene classes, including indoor/outdoor [38], [36], city/landscape [41], and sets of natural scenes (e.g. sunset/forest/mountain) [25]. However, as the number of categories increases, the issue of overlapping between scene classes in images arises. To handle this issue, a continuous organization of scene classes (e.g. from man-made to natural scenes) has been proposed [26]. Alternatively, the issue of scene class overlap can be addressed by doing scene annotation (e.g. labeling a scene as depicting multiple classes). This approach is followed by Boutell et al. [5], which exploits the output of one-against-all classifiers to derive multiple class labels. Although the attributions of multiple labels is not explored in our work, the framework we present, in particular the PLSA approach, can be easily extended to perform multi-label attribution [23].

Regarding global image representations for scene classification, the work by Vailaya et al. is regarded as representative of the literature in the field [41]. This approach relies on a combination of distinct low-level cues for different two-class problems (global edge features for city/landscape, and local color features for indoor/outdoor). In the work by Oliva and Torralba [26], an intermediate classification step into a

set of global image properties (*naturalness*, *openness*, *roughness*, *expansion*, and *ruggedness*) is proposed. Images are manually labeled with these properties, and a Discriminant Spectral Template (DST) is estimated for each property. The DSTs are based on the Discrete Fourrier Transform (DFT) extracted from the whole image, or from a four-by-four grid. A new image is represented by the degree of each of the five properties based on the corresponding estimated DST, and this representation is used for the classification into semantic scene categories (coast, country, forest, mountain, etc.). Other approaches to scene classification also rely on an intermediate supervised region classification step [25], [31], [8]. Based on a Bayesian Network formulation, Naphade and Huang defined a number of intermediate regional concepts (e.g. sky, water, rocks) in addition to the scene classes [25]. The relations between the regional and the global concepts are specified in the network structure. Serrano et al. [31] propose a two-stage classification of indoor/outdoor scenes, where features of individual image blocks from a spatial grid layout are first classified into indoor or outdoor. These local classification outputs are further combined to create the global scene representation used for the final image classification. Similarly, Vogel and Schiele recently used a spatial grid layout in a two-stage framework to perform scene retrieval and scene classification [42]. The first stage does classification of image blocks into a set of regional classes, which extends the set of classes defined in [25] (this requires block ground-truth labeling). The second stage performs retrieval or classification based on the occurrence of such regional concepts in query images. Alternatively, Lim and Jin [18] successfully used the soft output of semi-supervised regional concept detectors in an image indexing and retrieval application. In a different formulation, Kumar and Herbert used a conditional random field model to detect and localize man-made scene structures, doing in this way scene segmentation and classification [15]. Overall, a large number of local, regional, and global representations have been used for scene classification.

The combination of interest point detectors and local descriptors are increasingly popular for object detection, recognition, and classification [19]. The literature in the field is too large to discuss in details here [34], [12], [9], [7], [27], [35], [44], [17]. For the classification task, recent works include [12], [9], [7], [27], [10], [44]. Most existing works have targeted a relatively small number of object classes. Fergus et al. optimized, in a joint unsupervised model, a scale-invariant localized appearance model and a spatial distribution model [12]. Fei-Fei et al. proposed a method to learn object classes from a small number of training examples [9]. The same authors extended their work to an incremental learning procedure, and tested it on a large number of object categories [10]. Dorko and Schmid performed feature selection to identify local descriptors relevant to a particular object class, given weakly labeled training images [7]. Opelt et al. proposed to learn classifiers from a set of visual features, including local invariant ones, via boosting [27]. Although our work shares the use of invariant local descriptors with all these methods, scenes are different than objects in a number of ways, as discussed in the Introduction, and pose specific challenges.

The analogy between invariant local descriptors and words has also been exploited recently [34], [35], [44]. Sivic and Zisserman proposed to cluster and quantize local invariant features into visterms, for object matching in frames of a movie. Such approach allows to reduce noise sensitivity in matching and to search efficiently through a given video for frames containing the *same* visual content (e.g. an object) using inverted files [34], [35]. Willamowski et al. extended the use of visterms creating a system for object matching and classification based on a bag-of-words representation built from local invariant features and various classifiers [44]. However, these methods neither investigated the task of scene modeling and classification, nor considered latent aspect models as we do here.

In another research direction, a number of works have also relied on the definition of visterms and/or on variations of latent space models to model annotated images, i.e. to link images with key words [2], [4], [22], [45]. However, all these methods have relied on traditional regional image features without much viewpoint and/or illumination invariance. In our work, we characterize a scene using local descriptors as visterms, taking into account the problems that exist in the construction of a visterm vocabulary. We use latent space models not to annotate images but to address some limitations of the visterm vocabulary, describing images with a model that explicitly accounts for the importance of visterm co-occurrence.

In parallel to our work [29], [24], the joint use of local invariant descriptors and probabilitic latent aspect models has been investigated by Sivic et al. for object clustering in image collections [33], and by Fei-Fei and Perona for scene classification [11]. Although related, these two approaches differ from ours in their assumptions. Sivic et al. [33] investigated the use of both Latent Dirichlet Allocation (LDA) [3] and PLSA for clustering objects in image collections. With the same image representation as ours, they showed that

latent aspects closely correlate with object categories from the Caltech object data set, though these aspects are learned in an unsupervised manner. The number of aspects was chosen by hand to be equal (or very close) to the number of object categories, so that images are seen as mixtures of one 'background' aspect with one 'object' aspect. This allows for a direct match between object categories and aspects, but at the same time implies a strong coherence of the appearence of objects from the same category: each category is defined by only one multinomial distribution over the quantized local descriptors. Closer to our work, Fei-Fei and Perona [11] proposed two variations of LDA [3] to model scene categories. They tested different region detection processes to build an image representation based on quantized local descriptors. Contrarily to [33], Fei-Fei and Perona [11] propose to model a scene category as a mixture of aspects, and each aspect is defined by a multinomial distribution over the quantized local descriptors. This is achieved by the introduction of an observed class node in their models [11], which explicitly requires each image example to be labeled during the learning process.

In this paper, we model scene images using a probabilistic latent aspect model and quantized local descriptors, but without assuming a one-to-one correspondence between categories and aspects as in [33], and without learning a single distribution over aspects per scene category as in [11]. Images - not categories - are modeled as mixtures of aspects in a fully unsupervised way, without class information. The distribution over aspects serves as image representation, that is inferred on new images and used for supervised classification in a second step. These differences are crucial, as they allow us to investigate the use of unlabeled data for learning the aspect-based image representation.

III. Image representation

There are two main elements in an image classification system. The first one refers to the computation of the feature vector representing an image d, and the second one is the classifier, the algorithm that classifies an input image into one of the predefined category using the feature vector. In this section, we focus on the image representation and describe the two models that we use: the first one is the bag-of-visterms, built from quantized local descriptors, and the second one is obtained through the higher-level abstraction of the bag-of-visterms into a set of aspects using latent space modeling.

A. Bag-of-visterms representation from local descriptors

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

1) Interest point detection: The goal of the interest point detector is to automatically extract characteristic points -and more generally regions- from the image, which are invariant to some geometric and photometric transformations. This invariance property is interesting, as it ensures that given an image and its transformed version, the same image points will be extracted from both and hence, the same image representation will be obtained. Several interest point detectors exist in the literature. They vary mostly by the amount of invariance they theoretically ensure, the image property they exploit to achieve invariance, and the type of image structures they are designed to detect [40], [19], [21]. In this work, we use the difference of Gaussians (DOG) point detector [19]. This detector essentially identifies blob-like regions where a maximum or minimum of intensity occurs in the image, and it is invariant to translation, scale, rotation and constant illumination variations. We chose this detector since it has previously shown to perform well [20], and also since we found it to be a good choice in practice for the task at hand, performing competively compared to other detectors. The DOG detector is also faster and more compact than similarly performing detectors. An additional reason to prefer this detector over fully affine-invariant ones [21], [40], is also motivated by the fact that an increase of the degree of invariance may remove information about the local image content that is valuable for classification. An empirical evaluation of point detectors for classification will be presented in Section VII, see also Table IV.

2) Local descriptors: Local descriptors are computed on the region around each interest point identified by the local interest point detector. We use the SIFT (Scale Invariant Feature Transform) feature as local descriptors [19]. Our choice was motivated by findings in the literature [20], [11], where SIFT was found to work best; we also confirm this for our own work in Section VII. This descriptor is based on the grayscale representation of images. SIFT features are local histograms of edge directions computed over different parts of the interest region. These features capture the structure of the local image regions, which correspond to specific geometric configurations of edges or to more texture-like content. In [19], it was shown that the use of 8

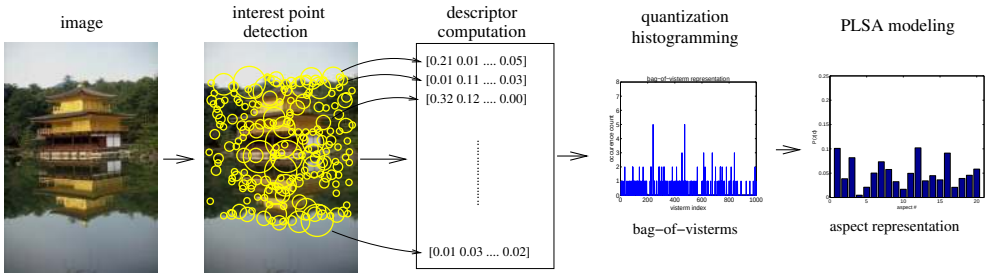

Fig. 1. Representation computation of an image.

orientation directions and a grid of 4x4 parts gives a good compromise between descriptor size and accuracy of representation. The size of the feature vector is thus 128. Orientation invariance is achieved by estimating the dominant orientation of the local image patch using the orientation histogram of the keypoint region. All direction computations to obtain the SIFT feature vector are done with respect to this dominant orientation.

3) Quantization and vocabulary model construction: When applying the two preceding steps to a given image, we obtain a set of real-valued local descriptors. 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:

$$s \longmapsto Q(s) = v_i \Longleftrightarrow \text{dist}(s, v_i) \leq \text{dist}(s, v_j), \qquad (1)$$

$\forall j \in \{1, \ldots, N_\mathcal{V}\}$, where $N_\mathcal{V}$ denotes the size of the visterm set. The set $\mathcal{V}$ of all visterms will be called vocabulary.

The construction of the vocabulary is performed through clustering. More specifically, we apply the K-means algorithm to a set of local descriptors extracted from training images, and the means are kept as visterms. We used the Euclidean distance in the clustering (and in Eq. 1) and choose the number of clusters depending on the desired vocabulary size. The choice of the Euclidean distance to compare SIFT features is common [19], [21].

Technically, the grouping of similar local descriptors into a specific visterm can be thought of as being similar to the *stemming* preprocessing step of text documents, which consists of replacing all words by their stem. The rationale behind stemming is that the meaning of words is carried by their stem rather than by their morphological variations [1]. The same motivation applies to the quantization of similar descriptors into a single visterm. Furthermore, in our framework, local descriptors will be considered as distinct whenever they are mapped to different visterms, regardless of whether they are close or not in the SIFT feature space. This also resembles the text modeling approach which considers that all information is in the stems.

4) Bag-of-visterms representation: The first representation of the image that we will use for classification is the bag-of-visterms (BOV), which is constructed from the local descriptors according to:

$$h(d) = (h_i(d))_{i=1..N_\mathcal{V}}, \text{ with } h_i(d) = \text{n}(d, v_i), \qquad (2)$$

where $\text{n}(d, v_i)$ denotes the number of occurrences of visterm v_i in image d. This vector-space representation of an image contains no information about spatial relationship between visterms. The standard bag-of-words text representation results in a very similar 'simplification' of the data: even though word ordering contains a significant amount of information about the original data, it is completely removed from the final document representation.

B. Probabilistic Latent Semantic Analysis (PLSA)

The bag-of-words approach has the advantage of producing a simple representation, but potentially introduces the well known *synonymy* and *polysemy* ambiguities, as will be shown in the next Section. Recently, probabilistic latent space models [14], [3] 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-occurrences can thus be considered using similar approaches, and we use the Probabilistic Latent Semantic Analysis [14] (PLSA) model in this paper for that purpose. Though PLSA suffers from a non-fully generative formulation, its tractable likelihood maximization makes it an interesting alternative to fully generative models [3] with comparative performance [33].

PLSA is a statistical model that associates a latent variable $z_l \in \mathcal{Z} = \{z_1, \ldots, z_{N_A}\}$, where $N_A$is the number of aspects, with each observation (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

$$P(v_j, d_i) = P(d_i) \sum_{l=1}^{N_A} P(z_l \mid d_i) P(v_j \mid z_l). \qquad (3)$$

PLSA introduces a conditional independence assumption, namely that the occurrence of a visterm v_j is independent of the image d_i it belongs to, given an aspect z_l. The model in Equation 3 is defined by the probability of an image $P(d_i)$, 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 aspect 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. More precisely, given a set of training images $\mathcal{D}$, the likelihood of the model parameters Θ can be expressed by

$$\mathcal{L}(\Theta|\mathcal{D}) = \prod_{d\in\mathcal{D}} \prod_{j=1}^{N_\mathcal{V}} p(v_j, d)^{\mathrm{n}(d,v_j)}, \qquad (4)$$

where the probability model is given by Eq. 3. The optimization is conducted using the Expectation-Maximization (EM) algorithm [14]. 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 representation of the image that we will use is defined by

$$a(d) = (P(z_l|d))_{l=1\ldots N_A}. \qquad (5)$$

IV. Analogy with text

In our framework, we consider the visterms like text terms and model them with techniques that are commonly applied to text. In this section, we compare properties of terms in documents with those of visterms within images. We first discuss the *sparsity* of the document representation, an important characteristic of text documents. We then consider issues related to the semantic of terms, namely *synonymy* and *polysemy*.

A. Representation sparsity

To investigate the analogy with text representation, we compare the behavior between the BOV representation of an image data set and the bag-of-words representation of a standard text categorization data set.

The REUTERS-21578[1] data set contains 12900 documents. The standard word stopping and stemming process produces a vocabulary of 17900 words. As previously observed in natural language statistics, the frequency of each word across the text data set follows the Zipf's law:$P_r = r^{-b}$, where r is the keyword rank according to its frequency and b is close to unity (see Fig. 2 (top)). This distribution results in an average number of 45 non-zero elements per document, which corresponds to an average sparseness of 0.25%. Out of the 17900 words in the dictionary, 35% occur once in the data set and 14% occur twice. Only 33% of the words appear in more than five documents.

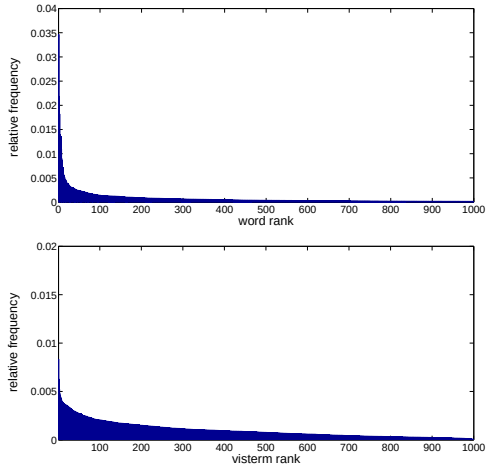

Fig. 2. Top: relative frequency distribution of the words extracted from REUTERS-21578, first 1000 words. Bottom: relative frequency distribution of the visterms in the city-landscape data set **D1**.

In our case, we applied the K-means algorithm on the **D1** image data set described in Section VI-B, which contains 6680 images of city and landscape, and generated the BOV representation for each image document of this data set for a vocabulary V_{1000} of size $N_v = 1000$. Since the visterm vocabulary is created by the K-means clustering of SIFT descriptors, the resulting vocabulary shows different properties than in text. As shown in Fig. 2 (right), the frequency distribution of visterms differs from the Zipf's law behavior usually observed in text. The K-means algorithm identifies regions in the feature space containing clusters of points, which prevents the low frequency effect observed in text data (see Fig. 2 bottom). The visterm with the lowest frequency appears in 117 images of the full data set (0.017 relative frequency). We also observed an average of 175 non-zero elements per image, which corresponds to a data sparseness of 17.5%.

The construction of the visual vocabulary by clustering intrinsically leads to a "flatter" distribution for visterms than for words. On one hand, this difference can be considered as an advantage, as the data sparseness observed in the text bag-of-words representation is indeed one of the main problems encountered in text retrieval and categorization. Similar documents might have very different bag-of-words representations because

[1] www.daviddlewis.com/resources/testcollections/reuters21578.

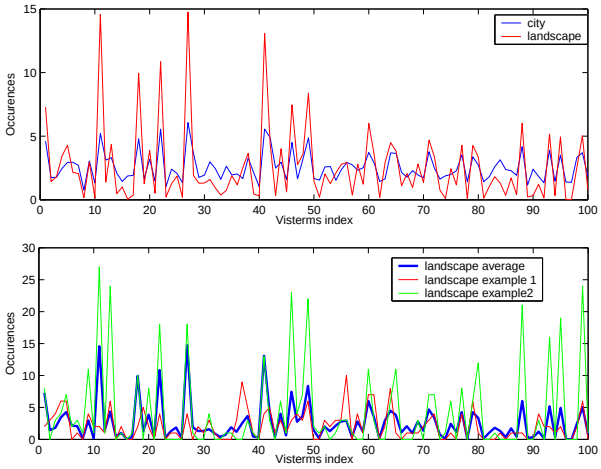

Fig. 3. Bag-of-visterms representation. Top: average of the BOV representation with respect to city (blue) and landscape (red) computed over the first split of data set **D1**. Bottom: landscape average (blue) compared with individual samples (red and green).

specific words in the vocabulary appear separately in their description. On the other hand, a flatter distribution of the features might imply that, on average, visterms in the visual vocabulary provide less discriminant information. In other words, the semantic content captured by individual visterms is not as specific as the one of words. We address this issue in the next subsection.

B. Polysemy and synonymy with visterms

To study the "semantic" nature of the visterms, we first considered the class conditional average of the BOV representation. Fig. 3 (top) shows the average of visterms for the city and landscape scene categories, computed over the first split of data set **D1** (see Section VI-B for details). We display the results when using the vocabulary of 100 visterms, V_{100}, defined in Section VII-A. The behavior is similar for other vocabulary sizes.

We first notice that there is a large majority of terms that appear in both classes: all the terms are substantially present in the city class; only a few of them do not appear in the landscape class. This contrasts with text documents, in which words are in general more specifically tied to a given category. Furthermore, we can also observe that the major peaks in the two class averages coincide in general. Thus, when using the BOV representation, the discriminant information with respect to the classification task seems to lie in the difference of average word occurrences. It is worth noticing that this is not due to a bias in the average in visterm numbers, since the difference in the average number of visterms per class is only in the order of 4% (city: 268/ landscape: 259). Additionally, these average curves hide the fact that there exists a large variability between samples, as illustrated in Fig. 3 (bottom), where two

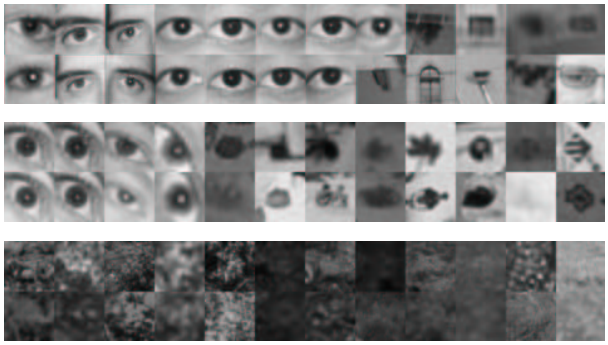

Fig. 4. Samples from three randomly selected visterms from a vocabulary of 1000 visterms.

random examples are plotted along with the average of the landscape class. Overall, all the above considerations indicate that visterms, taken in isolation, are not so class-specific, which in some sense advocates against feature selection only based on the analysis of the total occurrence of individual features (e.g. [7]), and reflects the fact that the semantic content carried by visterms, if any, is strongly subject to polysemy and synonymy issues.

To illustrate that visterms are subject to *polysemy* -a single visterm may represent different scene content- and *synonymy* -several visterms may characterize the same image content-, we show samples from three different visterms obtained when building the vocabulary V_{1000} (see Section VII-A for details) in Fig. 4. As can be seen, the top visterm (first two rows in Fig. 4) represents mostly eyes. However, windows and publicity patches get also indexed by this visterm, which provides an indication of the polysemic nature of that visterm, which means here that although this visterm will mostly occur on faces, it can also occur in city environments. The two middle rows in Fig. 4 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 of Fig. 4) indicate that this visterm captures a certain fine grain texture that has different origins (rock, trees, road or wall texture...), which illustrates that not all visterms have a clear semantic interpretation.

To conclude, it is interesting to notice that one factor that can affect the polysemy and synonymy issue is the vocabulary size: the polysemy of visterms might be more important when using a small vocabulary size than when using a large vocabulary. Conversely, with a large vocabulary, there are more chances to find many synonyms than with a small one. Since PLSA can in theory handle both synonymy and polysemy issues, it could in principle lead to a more stable representation for different vocabulary sizes.

V. Scene classification

To classify an input image d represented either by the BOV vectors h, the aspect parameters a, or any of the feature vector of the baseline approach (see next section), we employed Support Vector Machines (SVMs) [6]. SVMs have proven to be successful in solving machine learning problems in computer vision and text categorization applications, especially those involving large dimensional input spaces. In the current work, we used Gaussian kernel SVMs, whose bandwidth was chosen based on a 5-fold cross-validation procedure.

Standard SVMs are binary classifiers, which learn a decision function $f(x)$ through *margin* optimization [6], such that $f(x)$ is large (and positive) when the input x belongs to the target class, and negative otherwise. For multi-class classification, we adopt a one-against-all approach [43]. Given a n-class problem, we train n SVMs, where each SVM learns to differentiate images of one class from images of all other classes. In the testing phase, each test image is assigned to the class of the SVM that delivers the highest output of its decision function.

VI. Experimental setup

In this section, we describe the classification tasks we considered, the origin and composition of our data sets, the classification protocol we followed, and the baseline methods we used for comparison purposes.

A. Classification tasks

Four classification tasks, ranging from binary to five-class classification, have been considered to evaluate the performance of the proposed approaches. We first considered two standard, unambiguous binary classification tasks: indoor vs. outdoor, and landscape vs. city. These two tasks allow a first evaluation of the classification performance, and a fair comparison with approaches that have been proposed for the same tasks [41]. For a more detailed analysis of the performance, we then merged the two binary classification tasks to obtain a three-class problem (indoor vs. city vs. landscape). We also subdivided the landscape class into mountain and forest, and the city class into street view and panoramic view to obtain a five-class data set.

In Section IX we present additional results on two scene classification data sets, with 13 and 6 scene categories respectively, that have been proposed in recent literature [11], [42].

B. Datasets

Five data sets were created four our experiments: **D1:** this data set of 6680 images contains a subset of the Corel data set [41], and is composed of 2505 city and 4175 landscape images of 384×256 pixels. **D2:** this set is composed of 2777 indoor images retrieved from the Internet. The size of these images is typically 384×256 pixels. Original images with larger dimensions were resized using bilinear interpolation. The image size in the data set was kept approximately constant to avoid a potential bias in the BOV representation, since it is known that the number of detected interest points is highly dependent on the image resolution. **D3:** this data set is constituted by 3805 images from several sources: 1002 building images (ZuBud) [32], 144 people and outdoor images [27], 435 indoor human faces [44], 490 indoor images (Corel) [41], 1516 city/landscape overlap images (Corel) [41], and 267 Internet photographic images. **D4:** this data set is composed of all images from the data sets **D1** and **D2**. The total number of images in this data set is 9457. **D4v:** this is a subset of **D4** composed of 3805 randomly chosen images. **D5:** this is a five-class data set. It comprises all images from the data set **D2**, and images from **D1** whose content corresponds to the selected classes. From the 6680 images of **D1** we kept : 590 mountain images, 492 forest images, 1957 city street images (close-up of buildings), and 548 city panoramic images (middle to far views from buildings). The data sets contains a total of 6364 images.

In the experiments, We use the data set **D1** for the city vs. landscape scene classification task, and **D4** for indoor vs. outdoor scene classification, **D4** in the three-class case, and **D5** in the five-class problem.

Alternative vocabularies were constructed from either **D3** or **D4v**, allowing us to study the influence of the data on the vocabulary model, and its impact on classification performance. With 3805 images, we obtained in both cases approximately one million descriptors to train the vocabulary models. These data sets are available at: *http://carter.idiap.ch/data sets.html*.

C. Protocol

The protocol for each of the classification experiments was as follows. The full data set of a given experiment was divided into 10 parts, thus defining 10 different splits of the full data set. One split corresponds to keeping one part of the data for testing, while using the other nine parts for training (hence the amount of training data is 90% of the full data set). 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 parentheses after the mean value.

Additional experiments were conducted with less amount of training data, to test the robustness of the image representation. In that case, for each of the splits, images were chosen randomly from the training part of the split to create a reduced training set. Care was taken to keep the same class proportions in the reduced set as in the original set, and to use the same reduced training set in those experiments involving two different representation models. The test data of each split was left unchanged.

D. Baseline method

As a baseline method, we use the image representations proposed by Vailaya et al. [41]. We selected this approach, as it reports some of the best results from all scene classification approaches for data sets with landscape, city and indoor images on a significantly large data set. Thus, it can be regarded as a good representative of the state-of-the-art.

Two different representations are used for each binary classification tasks: color features are used to classify images as indoor or outdoor, and edge features are used to classify outdoor images as city or landscape. Color features are based on the LUV first- and second-order moments computed over a 10×10 spatial grid of the image, resulting in a 600-dimensional feature space. Edge features are based on edge coherence histograms calculated on the whole image, and are computed by extracting edges in only those neighborhoods exhibiting some edge direction coherence. Directions are then discretized into 72 directions, and their histogram is computed. An extra non-edge pixels bin is added to the histogram, leading to a feature space of 73 dimensions.

In the three-class problem Vailaya et al. apply both methods in a hierarchical way [41]. Images are first classified as indoor or outdoor given their color representation. All correctly classified outdoor images are further classified as either city or landscape, according to their edge direction histogram representation.

VII. Classification results

In this section, we present the classification results of our approach, first using the BOV representation, then using the aspect representation, and compare both of them with the baseline method. The performance of the methods under different conditions (vocabulary size, number of latent aspects, amount of training data) are presented and discussed.

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)
BOV V'_{100}	8.1	(0.5)	5.5	(0.9)
BOV V'_{300}	7.6	(0.9)	5.1	(1.2)
BOV V'_{600}	7.3	(0.8)	5.1	(0.7)
BOV V'_{1000}	7.2	(1.0)	5.4	(0.9)

TABLE I

Classification error for the baseline model and the BOV representation, for 8 vocabularies. Standard deviations are shown in parentheses.

A. Scene classification with bag-of-visterms

Binary classification

To analyze the effect of the size of the vocabulary 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, and constructed from **D3** as described in Section III. Additionally, four vocabularies V'_{100}, V'_{300}, V'_{600}, and V'_{1000} were constructed from **D4v**.

Table I provides the classification error for the two binary classification tasks. We can observe that the BOV approach consistently outperforms the baseline methods. This is confirmed in all cases by a paired T-test, for $p = 0.05$. It is important to remind that contrarily to the baseline methods, the BOV representation uses the same features for both tasks and no color information.

Regarding vocabulary size, overall we can see that for vocabularies of 300 visterms or more the classification errors are equivalent. This contrasts with the work in [44], where the 'flattening' of the classification performance was observed only for vocabularies of 1000 visterms or more. A possible explanation may come from the difference in task (object classification) and in the use of the Harris-Affine point detector [21], known to be less stable than DOG [20].

The comparison of the rows 2-5 and 6-9 in Table I shows that using a vocabulary constructed from a data set *different* than the one used for the classification experiments, **D3** and **D4v** respectively, does not affect the results (error rates differences are within random fluctuation values). This result confirms the observations made in [44], and suggests that it might be feasible to build a generic visterm vocabulary that can be used for different tasks. Based on these results, we use the vocabularies built from **D3** in all the remaining experiments.

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 II

THREE-CLASS CLASSIFICATION ERROR FOR BASELINE AND BOV MODELS. THE BASELINE MODEL SYSTEM IS HIERARCHICAL.

Total class. error				11.1 (0.8)	
	Classification (%)			Class.	# of
Gr. Truth	indoor	city	land.	Error (%)	images
indoor	89.7	9.0	1.3	10.3	2777
city	14.5	74.8	10.7	25.2	2505
landscape	1.2	2.0	96.8	3.1	4175

TABLE III

CONFUSION MATRIX FOR THE THREE-CLASS CLASSIFICATION PROBLEM, USING VOCABULARY V_{1000}.

Three-class classification

Table II shows the results of the BOV approach for the three-class classification problem. 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 approach proposed in [41] with statistically significant differences. This is confirmed in all cases by a paired T-test, with $p = 0.05$. Secondly, we observe the stability of results with vocabularies of 300 or more visterms, the vocabulary of 1000 visterms giving slightly better performance. Based on these results, we assume V_{1000} to be an adequate choice and use V_{1000} for all experiments in the rest of this paper. Finally, we can observe that the classification strategy, hierarchical or multi-class SVM, has little impact on the results for this task.

A closer analysis of the results can be done by looking at the confusion matrix, shown in Table III. First, we can see that landscape images are well classified. Secondly, we observe that there exists some confusion between the indoor and city classes. This can be explained by the fact that both classes share not only similar local image structures (which will be reflected in the same visterms appearing in both cases), but also similar visterm distributions, due to the resemblance between some more general patterns (e.g. doors or windows). The two images on the top in Fig. 5 illustrate some typical errors made in this case, when city images contain a majority of geometric shapes and little texture. In the third place, the confusion matrix also shows that city images are also misclassified as landscape. The main explanation is that city images often contain natural elements (vegetation like trees or flowers, or natural textures), and specific structures which produce many visterms. The images to the bottom in Fig. 5 illustrate typical mistakes in this case.

Fig. 5. Typical classification errors of city images in the three-class problem. Top: city images classified as indoor. Bottom: city images classified as landscape.

We now explore different combinations of point detectors/descriptors. We purposely choose to do this study on the 3-class problem since we believe that a multi-class classification task is a more representative problem for this data, but at the same time it is not obscured by many of the additional issues of a many-class task. Four point detection methods: DOG [19], multi-scale Harris affine (MHA) [21], multi-scale Harris (MH) [21], and a fixed 15x20 grid (GRID), and three descriptor methods: SIFT [19], complex filters (CF) [30], and a 11×11 pixel sample of the area defined by the detector (PATCH) were used in paired combinations. The results are shown in Table IV.

	SIFT	CF	PATCH	av. # of points
DOG	11.1 (0.8)	22.5 (1.1)	22.1 (0.9)	271
MHA	11.9 (1.1)	18.4 (1.1)	20.6 (1.3)	424
MH	11.8 (1.0)	19.3 (0.9)	-	580
GRID	19.9 (0.9)	-	19.8 (0.8)	300

TABLE IV

COMPARISON OF COMBINATIONS OF DETECTOR/DESCRIPTORS FOR INDOOR/CITY/LANDSCAPE CLASSIFICATION. THE AVERAGE NUMBER OF DETECTED POINTS PER IMAGE IS ALSO SHOWN.

In Table IV, we can see that the combination DOG+SIFT is the best performing one, this is confirmed

Total class. error rate: 20.8 (2.1) (Baseline: 30.1 (1.1))							
	m.	f.	i.	c.-p.	c.-s.	error (%)	# of images
mount.	85.8	8.6	2.5	0.5	2.6	14.2	590
forest	8.9	80.3	1.6	2.4	6.7	19.7	492
indoor	0.4	0	91.1	0.4	8.1	8.9	2777
city-pan.	3.5	1.8	8.0	46.9	39.8	53.1	549
city-str.	2.0	2.2	20.8	6.0	68.9	31.1	1957

TABLE V
CLASSIFICATION RATE AND CONFUSION MATRIX FOR THE FIVE-CLASS, USING BOV AND VOCABULARY V_{1000}.

by a paired T-test, with $p = 0.05$. However, MHA+SIFT and MH+SIFT produce similar results. This confirms SIFT as the best performing descriptor, as pointed out in the literature, although for other tasks [11], [20]. As for detectors, it is important to note that, although the multi-scale Harris and multi-scale Harris affine detectors [21] allow for similar performance, DOG is computationally more efficient and more compact (less feature points per image). Although Table IV shows DOG+SIFT to be the best choice for this particular task, it is possible that other combinations may perform better for other tasks. Based on these results, however, we have confirmed in practice that DOG+SIFT constitutes a reasonable choice.

Five-class classification

Table V presents the overall error rate and the confusion matrix obtained with the BOV approach in the five-class experiment, along with the baseline overall error rate. The latter number was obtained using the edge coherence histogram global feature [41].

The BOV representation performs much better than the global features in this task, and the results show that we can apply the BOV approach to a larger number of scene classes and obtain good results.

Analyzing the confusion matrix, we first observe that some mistakes are made between the forest and mountain classes, reflecting their sharing of similar textures and the presence of forest in some mountain images. A second observation is that city-panorama images are often confused with city-street images. This result is not surprising because of the somewhat ambiguous definition of the classes (see Fig. 6), which was already observed during the human annotation process. The errors can be further explained by the scale-invariant nature of the interest point detector, which makes no distinction between some far-field street views in the city-panoramic images, and middle-view similar structures in the city-street images. Another explanation is the unbalanced data set, with almost four times as many city-street images than panoramic ones. Finally, we observe that the main source of confusion lays between the indoor images and the city-street images, for similar reasons as those described in the three-class task.

Fig. 6. Illustration of the five classes, with 8 randomly selected examples per class. From left to right: mountain, forest, indoor, city-panorama, city-street. All images have been cropped for display.

B. Scene classification with PLSA

In PLSA, we use the probability distribution $P(z_l|d_i)$ of latent aspects given each specific document as a N_A dimensional feature vector $a(d)$ (Eq. 5). Given that PLSA is an unsupervised approach, where no reference to the class label is used during the aspect model learning, we may wonder how much discriminant information remains in the aspect representation. To answer this question, we compare the classification errors obtained with the PLSA and BOV representations. Furthermore, to test the influence of the training data on the aspect model, we conducted two experiments which only differ in the data used to estimate the $P(v_j|z_l)$ multinomial probabilities. More precisely, we defined two cases:

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

PLSA-O: the aspect models are trained only once on the auxiliary data set **D3**, which is disjoint from the sets used for SVM learning.

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

Method	A	ind./out.	city/land.	ind./city/land.
BOV		7.6 (1.0)	5.3 (1.1)	11.1 (0.8)
PLSA-I	20	9.5 (1.0)	5.5 (0.9)	12.6 (0.8)
PLSA-I	60	8.3 (0.8)	4.7 (0.9)	11.2 (1.3)
PLSA-O	20	8.9 (1.4)	5.6 (0.9)	12.3 (1.2)
PLSA-O	60	7.8 (1.2)	4.9 (0.9)	11.9 (1.0)

TABLE VI

COMPARISON OF BOV, PLSA-I AND PLSA-O STRATEGIES ON THE TWO- AND THREE-CLASS CLASSIFICATION TASKS, USING 20 AND 60 ASPECTS, AND VOCABULARY V_{1000}.

N_A	20	40	60	80	100
Error	5.6 (0.9)	4.9 (0.8)	4.9 (0.9)	4.8 (1.0)	5.0 (0.9)

TABLE VII

CLASSIFICATION RESULTS FOR THE CITY/LANDSCAPE TASK, USING DIFFERENT NUMBER OF ASPECTS FOR PLSA-O.

Classification results: two and three-class cases

Table VI shows the classification performance of the latent space representation for 20 and 60 aspects for the two strategies PLSA-I and PLSA-O, using V_{1000}. The corresponding results for BOV with the same vocabulary are re-displayed for comparison purposes.

Discussing first the PLSA training data issue, we observe that performance of both strategies is comparable for the city/landscape scene classification, being PLSA-O better than PLSA-I for indoor/outdoor (paired T-test, with $p = 0.05$). This might suggest that aspect models learned on the same set used for SVM training may cause some over-fitting in the indoor/outdoor case. Since using PLSA-O allows to learn one single model for all tasks, we chose this approach for the rest of the experiments. Of course, the data set from which the aspects are learned must be sufficiently representative of the collection to be classified in order to obtain a valid aspect-based representation.

Comparing the 60-aspect PLSA-O model with the BOV approach, we observe that their performance is similar, and that PLSA performs better in the city/landscape case (although not significantly), while the opposite holds for the three-class task. Learning visual co-occurrences with 60 aspects in PLSA allows for dimensionality reduction by a factor of 17 while keeping the discriminant information contained in the original BOV representation. Note that PLSA with 60 aspects performs better than the BOV representation with the vocabulary V_{100} in all cases (see Tables I and II).

We also conducted experiments to study the importance of the number of aspects on the classification performance. Table VII displays the evolution of the error with the number of aspects for the city/landscape classification task. The results show that the performance is relatively independent of the number of aspects in the range [40,100]. For the rest of this paper we use a PLSA model with $N_A = 60$ aspects.

For comparison purposes, we present in Table VIII the confusion matrix in the three-class classification task. The errors are similar to those obtained with the BOV (Table III). The only noticeable difference is that more indoor images were misclassified in the city class.

Decreasing the amount of training data

Since PLSA captures co-occurrence information from the data it is learned from, it can provide a more stable image representation. We expect this to help in the case of lack of sufficient labeled training data for the classifier. Table IX compares classification errors for the BOV and the PLSA representations for the different tasks when using less data to train the SVMs. The amount of training data is given both in proportion to the full data set size, and as the total number of training images. The test sets remain identical in all cases.

Several comments can be made from this table. A general one is that for all methods, the larger the training set, the better the results, showing the need for building large and representative data sets for training purposes. Qualitatively, with the PLSA and BOV approaches, performance degrades smoothly initially, and sharply when using 1% of training data. With the baseline, on the other hand, performance degrades more steadily.

Comparing methods, we first notice that PLSA with 10% of training data outperforms the baseline approach with full training set (i.e. 90%), this is confirmed in all cases by a Paired T-test, with $p = 0.05$. BOV with 10% of training still outperforms the baseline approach with full training set (i.e. 90%) for indoor/outdoor (paired T-test with $p = 0.05$). More generally, we observe that both PLSA and BOV perform better than the baseline for -almost- all cases of reduced training set. An exception is the city/landscape classification case, where the baseline

Total class. error				11.9(1.0)	
	indoor	city	land.	class error(%)	# images
indoor	86.6	11.8	1.6	13.4	2777
city	14.8	75.4	9.8	24.5	2505
land.	1.3	1.9	96.8	3.1	4175

TABLE VIII

CLASSIFICATION ERROR AND CONFUSION MATRIX FOR THE THREE-CLASS PROBLEM USING PLSA, WITH V_{1000} AND 60 ASPECTS.

Method	Amount of training data				
	90%	10%	5%	2.5%	1%
Indoor/Outdoor					
# images	8511	945	472	236	90
PLSA	7.8(1.2)	9.1(1.3)	10.0(1.2)	11.4(1.1)	13.9(1.0)
BOV	7.6(1.0)	9.7(1.4)	10.4(0.9)	12.2(1.0)	14.3(2.4)
Baseline	10.4(0.8)	15.9(0.4)	19.0(1.4)	23.0(1.9)	26.0(1.9)
City/Landscape					
# images	6012	668	334	167	67
PLSA	4.9(0.9)	5.8(0.9)	6.6(0.8)	8.1(0.9)	17.1(1.2)
BOV	5.3(1.1)	7.4(0.9)	8.6(1.0)	12.4(0.9)	30.8(1.1)
Baseline	8.3(1.5)	9.5(0.8)	10.0(1.1)	11.5(0.9)	13.9(1.3)
Indoor/City/Landscape					
# images	8511	945	472	236	90
PLSA	11.9(1.0)	14.6(1.1)	15.1(1.4)	16.7(1.8)	22.5(4.5)
BOV	11.1(0.8)	15.4(1.1)	16.6(1.3)	20.7(1.3)	31.7(3.4)
Baseline	15.9(1.0)	19.7(1.4)	24.1(1.4)	29.0(1.6)	33.9(2.1)

TABLE IX

CLASSIFICATION PERFORMANCE FOR PLSA-O WITH 60 ASPECTS, BOV WITH VOCABULARY V_{1000}, AND BASELINE APPROACHES, WHEN USING A SVM CLASSIFIER TRAINED WITH PROGRESSIVELY LESS DATA. THE AMOUNT OF TRAINING DATA IS GIVEN AS PERCENTAGE OF THE FULL DATA SET, AND THEN AS THE ACTUAL NUMBER OF TRAINING IMAGES.

is better than the BOV when using 2.5% and 1% training data, and better than the PLSA model for 1%. This can be explained by the fact that edge orientation features are particularly well adapted for this task, and that with only 25 city and 42 landscape images for training, global features are competitive.

Furthermore, we notice that PLSA deteriorates less as the training set is reduced, producing better results than BOV for all reduced training set experiments (although not always significantly better).

Previous work on probabilistic latent space modeling has reported similar behavior for text data [3]. PLSA's better performance in this case is likely due to its ability to capture aspects that contain general information about visual co-occurrence. Thus, while the lack of data impairs the simple BOV representation in covering the space of documents belonging to a specific scene class (eg. due to the synonymy and polysemy issues) the PLSA-based representation is less affected.

Classification results: five-class case

Table X reports the overall error rate and the confusion matrix obtained with PLSA-O in the five-class problem, and with the full training set. As can be seen, PLSA performs slightly worse than BOV, but still better than the baseline. By comparing the confusion matrix with that of the BOV case (Table V), we can see that, while the forest, mountain, and indoor classification

Total error rate (BOV: 20.8 (2.1), Baseline: 30.1 (1.1))						
	m.	f.	i.	c.-p.	c.-s.	error (%)
mountain	85.5	12.2	0.8	0.3	1.2	14.5
forest	12.8	78.3	0.8	0.4	7.7	21.7
indoor	0.3	0.1	88.9	0.2	10.5	11.1
city-pan.	3.6	4.9	8.8	12.6	70.1	87.4
city-str.	1.6	1.4	20.4	1.7	74.9	25.1

TABLE X

CLASSIFICATION ERROR AND CONFUSION MATRIX FOR THE FIVE-CLASS PROBLEM USING PLSA-O WITH 60 ASPECTS.

Perc. data	90%	10%	5%	2.5%	1%
# images	5727	636	318	159	64
PLSA	23.1(1.2)	27.9(2.2)	29.7(2.0)	33.1(2.5)	38.5(2.6)
BOV	20.8(2.1)	25.5(1.7)	28.3(1.3)	30.8(1.6)	37.2(3.4)
Baseline	30.1(1.1)	36.8(1.4)	39.3(1.4)	42.8(1.6)	49.9(3)

TABLE XI

COMPARISON BETWEEN BOV, PLSA-O, AND BASELINE, FOR SVM TRAINED WITH REDUCED DATA ON THE 5-CLASS PROBLEM.

behavior remains almost unchanged, the results for the two city classes were significantly altered. The main explanation comes from the rather loose definition of the city-panorama class, which contains many more images from landmark buildings in the middle distance than 'cityscape' images. Due to this fact, combined with the visterm scale invariance, the PLSA modeling generates a representation for the city-panorama images which clearly contains building-related aspects, and introduces confusion with the city-street class. In this case, the abstraction level of PLSA loses some of the discriminative elements of the BOV. Due to the unbalanced data set, the city-street class beneficiates from this confusion, as shown by its reduced misclassification rate with respect to the city-panorama class. Furthermore, aspects are learned on the **D3** data set, which contains a relatively small amount of city-panorama images compared to city-street images. This imbalance can explain the ambiguous aspect representation of the city-panorama class and the resulting poor classification performance.

Table XI presents the evolution of the classification error when less labeled training data is available. It shows that the loss of discriminative power between the city-panorama and city-street classes continue to affect the PLSA representation, and that, in this task, the BOV approach outperforms the PLSA model for reduced training data. Both methods, however, perform better than the global approach.

The five-class experiment raises a more general issue. As we introduce more classes or labels, the possibility of defining clear-cut scenes and of finding images that belong to only one class diminishes, while the number

of images whose content belongs to several concepts increases. With more classes, the task could be better formulated as an annotation problem rather than a classification one. PLSA-based approaches have shown promising performance for this task [23].

In the case of less confusing class definitions, the PLSA approach can be valid for other multi-class problems. We have recently applied our approach on a seven-class object data set with good performance (88% classification rate), and obtaining similar conclusions with respect to the properties of our approach [24].

We have performed additional experiments with more classes on Section IX where we investigate the application of both BOV and PLSA scene modeling to problems with more classes (13 and 6).

VIII. Aspect-based image ranking

With PLSA, aspects can be conveniently illustrated by their most probable images in a data set. Given an aspect z, images can be ranked according to:

$$P(d|z) = \frac{P(z|d)P(d)}{P(z)} \propto P(z \mid d), \qquad (6)$$

where $P(d)$ is considered as uniform. The top-ranked images for a given aspect illustrate its potential 'visual meaning'. Fig. 7 displays the 10 most probable images from the 668 test images of the first split of the **D1** data set, for seven out of 20 aspects learned on the **D3** data set. The top-ranked images representing aspects 1, 6, 8, and 16 all clearly belong to the landscape class. More precisely, aspect 1 seems to be mainly related to horizon/panoramic scenes, aspect 6 and 8 to forest/vegetation, and aspect 16 to rocks. Conversely, aspect 4 and 12 are related to the city class. However, as aspects are identified by analyzing the co-occurrence of local visual patterns, they 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. The results can be better appreciated at *http://carter.idiap.ch/aspect_ranking/index.html*.

Considering the aspect-based image ranking as an information retrieval system, the correspondence between aspects and scene classes can be measured objectively. 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 the precision/recall curves associated with each aspect-based image ranking considering either city and landscape queries, as illustrated in Fig. 8. Those

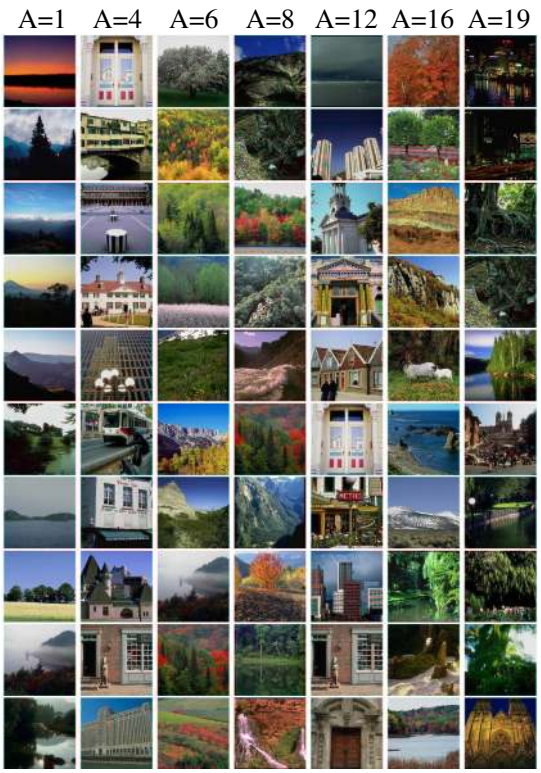

Fig. 7. The 10 most probable images from the **D1** data set for seven aspects (out of 20) learned on the **D3** data set.

curves prove that some aspects are clearly related to such concepts, and confirm observations made previously with respect to aspects 4, 6, 8, 12, and 16. As expected, aspect 19 does not appear in either the city or landscape top precision/recall curves. The landscape-related ranking from aspect 1 does not hold as clearly for higher recall values, because the co-occurrences of the visterm patterns appearing in horizons that it captures is not exclusive to the landscape class. Overall, these results illustrate that the latent structure identified by PLSA highly correlates with the visual structure of our data. This potentially makes PLSA a very attractive tool for browsing/annotating unlabeled image collections.

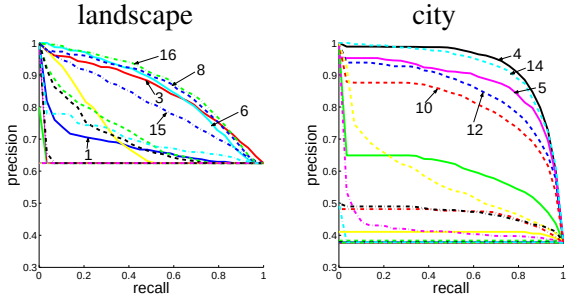

Fig. 8. Precision/recall curves for the image ranking based on each of the 20 individual aspects, relative to the landscape (left) and city (right) query. Each curve represents a diferent aspect. Floor precision values correspond to the proportion of landscape(resp. city) images in the data set.

IX. Experiments with other data sets

Given the recent appearance of other works and data sets in works on scene classification [11], [42], we have also compared our framework to them. In [11], the authors tackle the classification of 13 different scene types. In [42], the authors tackle the classification of 6 different natural scenes types, all collected from outdoor images. We present a short description of those data sets in the next paragraphs.

13-class data set [11] This data set contains a total of 3859 images of approx. 60000 pixel resolution, varying in exact size and XY ratio. The images are distributed over 13 scene classes as follows (the number in parenthesis indicates the number of images in each class): bedroom (216), coast (360), forest (328), highway (260), inside city (308), kitchen (210), living room (289), mountain (374), open country (410), office (215), street (292), suburb (241), and tall buildings (356) (available for download at: http://faculty.ece.uiuc.edu/feifeili/data sets.html).

6-class data set [42] This relatively small data set contains a total of 700 images of resolution 720×480 pixels. They are distributed over 6 natural scene classes as follows: coasts (142), river/lakes (111), forests (103), plains (131), mountains (179), and sky/clouds (34).

These two data sets are challenging given their respective number of classes and the intrinsic ambiguities that arise from their definition. In the 13-class data set for example, images from the inside city and street categories share a very similar scene configuration. Similarly, the differences between bedroom and living room examples can be subtle. In the 6-class data set, examples of the coasts and waterscapes classes are hard to distinguish. The same ambiguous class definition was observed for our five-class classification task in Section VII-A.

In Section VII, we evaluated visterm vocabularies built from different data sources, and conducted a comparison of aspect representations learned from extra data (PLSA-O) or learned on the same data used to learn the SVM classifier (PLSA-I). Given that we have no extra set of representative images for the 13-class or 6-class data, we can not present the same experiments for these data sets. To keep consistency with the way in which results are presented in [11], [42], we report classification accuracy instead of classification error.

A. Classification results: 13-class

We first classify the images based on their BOV as in Section VII. Results were obtained by training a multi-class SVM using a 10-split protocol, as in Section VI-C. No parameter tuning on the vocabulary was done in this case, as we directly apply the vocabulary V_{1000} used in Section VII.

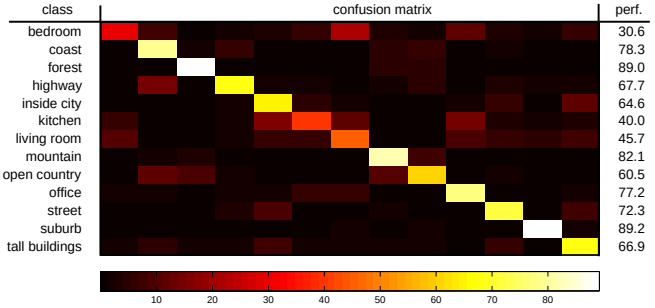

Fig. 9. Classification accuracy for the BOV representation, in the 13-class problem from [11]. The overall classification accuracy is 66.5%.

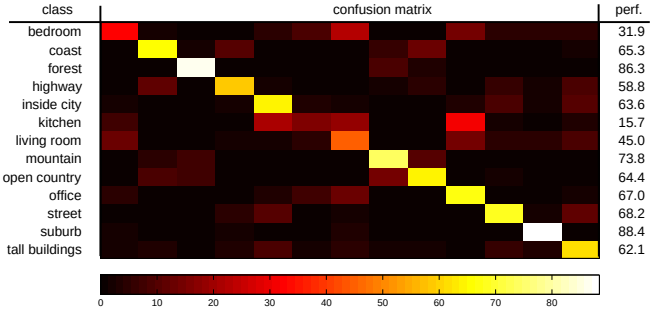

Fig. 10. Classification accuracy for the PLSA-I representation, in the 13-class problem from [11]. The overall classification accuracy is 60.8%.

The confusion matrix for the 13 classes and the classification performance per class are presented in Figure 9. The classification performance is substantially higher than the one presented by [11], which reported an overall classification performance of 52.5% when using the same combination of detector/descriptors we adopted here (DOG+SIFT) for learning their model. The performance of our method is also slightly better than the *best* performance reported in [11] (65.2%, obtained with a different detector/descriptor pair: GRID/SIFT). As we do not have access to the individual per-image results of [11], we cannot assess the statistical significance of these results, but we can nevertheless consider that the BOV approach is competitive.

We also applied the PLSA-I approach to solve the same classification problem, as in Section VII-B. We learned PLSA with 40 aspects, since this is the number of aspects used in [11]. Results were obtained, as before, with a multi-class SVM trained using a 10-split protocol.

Figure 10 shows the performance of the PLSA-I representation. The classification accuracy is higher than the one in [11] when using the (DOG+SIFT) combination, but is lower than the *best* performance reported

Class	confusion matrix						perf.
coasts	59.9	9.9	2.1	8.5	18.3	1.4	59.8
river/lakes	1.6	24.3	10.8	10.8	27.0	5.4	24.3
forests	2.9	5.8	81.6	4.9	4.9	0.0	81.6
plains	18.3	6.1	8.4	52.7	11.5	3.1	52.7
mountains	11.2	8.9	2.2	2.8	73.7	1.1	73.7
sky/clouds	5.9	2.9	0.0	5.9	5.9	79.4	79.4
overall							61.9

TABLE XII
CLASSIFICATION ACCURACY FOR THE BOV REPRESENTATION, IN THE 6-CLASS PROBLEM PRESENTED IN [42].

Class	confusion matrix						perf.
coasts	40.1	9.9	9.2	12.0	25.4	3.5	40.1
river/lakes	20.7	21.6	11.7	12.6	30.6	2.7	21.6
forests	1.9	3.9	78.6	7.8	7.8	0.0	78.6
plains	20.6	6.9	11.5	35.9	21.4	3.8	35.9
mountains	8.4	7.3	11.7	5.6	65.9	1.1	65.9
sky/clouds	14.7	0.0	0.0	8.8	5.9	70.6	70.6
overall							52.1

TABLE XIII
CLASSIFICATION FOR THE PLSA-O REPRESENTATION, IN THE 6-CLASS PROBLEM PRESENTED IN [42].

in [11], and also lower that one obtained with BOV. The performance degradation between BOV and PLSA results from the same phenomena observed for the five-class experiments in Section VII-B. In the presence of a high number of classes, the PLSA decomposition tends to result in a loss of important details for the distinction of ambiguous classes. As with the BOV case, we can also say that the PLSA approach remains competitive with respect to [11].

B. Classification results: 6-class

The data set presented by Vogel et al. [42] is composed of less classes than [11], with a total of six natural scene types. The ambiguity between class definitions is however more important, and some images are difficult to classify in only one scene type. The number of examples per class is significantly smaller than that in [11] and than the five-class data set in Section VII.

The multi-class SVM results, obtained using a 10-split protocol on the BOV representations (V_{1000} vocabulary learned on **D3**) are presented in Table XII. In this case, our system has a slightly reduced classification accuracy (61.9%) when compared with the performance presented in [42](67.2%). Note, however, that these results have not been obtained using identical features: [42] relies on a fixed grid, where a texture and color features are extracted. We believe that the difference in performance with respect to our work arises from the fact that natural scene discrimination can benefit greatly from the use of color, something we have not made use of, but which in light of these results constitutes an issue to investigate in the future. Moreover, the intermediate classification step proposed in [42] requires the expensive manual labeling of hundreds of regional descriptors, which is not needed in our case.

Given the reduced set of examples per class, and the need for a large number of representative examples to train a PLSA model, we could not perform the PLSA-I approach for this 6-class problem. However, in order to evaluate the performance of the aspect representation for these data, we use the previous PLSA model with 60 aspects learned on the **D3** data set (see Section VII-B). The corresponding classification results, as shown in Table XIII, indicate a decrease in performance (52.1%) with respect to both BOV and the results reported in [42]. The fact that the PLSA model has been learned on the **D3** data set, which does not contain any coasts, river/lakes, or plain examples, likely explains the poor discrimination between the 6-classes when the aspect representation is used.

Overall, these experiments support some of the findings obtained in Section VII, namely that modeling scenes as a *bag-of-visterms* performs well even in problems with a large amount of classes, and that PLSA modeling can find limitations in cases of large amount of overlapping classes. At the same time, these experiments offer other insights: our framework is competitive with recent approaches, and feature fusion mechanisms (adding color) have a potential for an increased classification performance.

X. CONCLUSION

Based on the results presented in this paper, we believe that the presented scene modeling methodology is effective for solving scene classification problems. We have shown, with extensive results, that it outperforms classical scene classification methods. We have also shown that it is able to handle a variety of problems without having to redesign the features used.

Regarding the specific contributions of this paper, we first presented results that demonstrate that the *bag-of-visterms* approach is adequate for scene classification, consistently outperforming methods relying on a suite of hand-picked global features. In the second place, we also showed that the PLSA-based representation is competitive with the BOV in terms of performance and results,

in general, in a more graceful performance degradation with decreasing amount of training data. This result is potentially relevant for the portability and re-usability of future systems, since it allows to reuse a classification system for a new problem using less training data. Thirdly, we also demonstrated that PLSA-based clustering of images reveals visually coherent grouping that we showed to be valuable for aspect-based image ranking. Finally, as part of our work, we explored the visterm vocabulary co-occurrence properties, and compared them to those of words in text documents. The results of such analysis showed the presence of cases of synonymy and polysemy as in text words, but also showed other statistical properties, such as sparsity, to be different than those in text. This, we believe, is mainly due to the vocabulary construction methodology, and advocates for improved vocabulary construction approaches.

The description of a visual scene as a mixture of aspects is an intriguing concept worth of further exploration. We are currently exploring the extension of PLSA modeling for scene segmentation. Further areas to investigate with the approach are the extraction of more meaningful vocabularies, the study of the influence of the degree of invariance of the local descriptors, and the definition of feature fusion mechanisms (e.g. color and local descriptors) in the latent space framework.

ACKNOWLEDGMENTS

This work was partially funded by the European Network of Excellence "PASCAL", through the project "CARTER", and by the Swiss NCCR (IM)2. T. Tuytelaars is supported by the Fund for Scientific Research Flanders. We thank Mihai Osian for discussions.

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.

[2] 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.

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

[4] D. Blei and M. Jordan. Modeling annotated data. In *Proc. 26th Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, Toronto, Aug. 2003.

[5] M.R. Boutell, J. Luo, X. Shen, and C.M.Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, Sep. 2004.

[6] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[7] G. Dorko and C. Schmid. Selection of scale invariant parts for object class recognition. In *Proc. of IEEE Int. Conf. on Computer Vision*, Nice, Oct. 2003.

[8] J. Fauqueur and N. Boujemaa. New image retrieval paradigm: logical composition of region categories. In *Proc. Int. Conf. on Image Processing*, Barcelona, Spain, October 2003.

[9] L. Fei-Fei, R. Fergus, and P. Perona. A Bayesian approach to unsupervised one-shot learning of object categories. In *Proc. of IEEE Int. Conf. on Computer Vision*, Nice, Oct. 2003.

[10] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *Proc. of IEEE Int. Conf. on Computer Vision, Workshop on Generative-Model Based Vision*, Washington DC, Jun. 2004.

[11] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proc. of. IEEE Int. Conf. on Computer Vision And Pattern Recognition*, San Diego, Jun. 2005.

[12] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Toronto, Jun. 2003.

[13] M. Gorkani and R. Picard. Texture orientation for sorting photos at glance. In *Proc. of Int. Conf. on Pattern Recognition*, Jerusalem, Sep. 1994.

[14] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.

[15] S. Kumar and M. Herbert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Proc. of IEEE Int. Conf. on Computer Vision*, Nice, Oct. 2003.

[16] S. Kumar and M. Herbert. Man-made structure detection in natural images using a causal multiscale random field. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Toronto, Jun. 2003.

[17] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *Proc. of the British Machine Vision Conference*, Norwich, Sep. 2003.

[18] J.-H. Lim and J.S. Jin. Semantics discovery for image indexing. In *European Conference on Computer Vision ECCV'04*, Prague, Czech Republic, May 2004.

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[20] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Toronto, Jun. 2003.

[21] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65:43–72, 2005.

[22] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proc. ACM Int. Conf. on Multimedia*, Berkeley, Nov. 2003.

[23] F. Monay and D. Gatica-Perez. PLSA-based image auto-annotation: Constraining the latent space. In *Proc. ACM Int. Conf. on Multimedia*, New York, Oct. 2004.

[24] F. Monay, P. Quelhas, D. Gatica-Perez, and J.-M. Odobez. Constructing visual models with a latent space approach. In *Proc. of the PASCAL Workshop on Subspace, Latent Structure and Feature Selection techniques: Statistical and Optimisation perspectives*, Bohinj, Feb. 2005.

[25] M. Naphade and T. Huang. A probabilistic framework for semantic video indexing, filtering and retrieval. *IEEE Trans. on Multimedia*, 3(1):141–151, Mar. 2001.

[26] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

[27] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proc. of IEEE Europ. Conf. on Computer Vision*, Prague, May 2004.

[28] S. Paek and Chang. S.-F. A knowledge engineering approach for image classification based on probabilistic reasoning systems. In *Proc. of IEEE Int. Conference on Multimedia and Expo*, New York, Aug. 2000.

[29] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *Proc. of IEEE Int. Conf. on Computer Vision*, Beijing, Oct. 2005.

[30] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets. In *European Conference on Computer Vision ECCV'02*, 2002.

[31] N. Serrano, A. Savakis, and J. Luo. A computationally efficent approach to indoor/outdoor scene classification. In *International Conference on Pattern Recognition*, Quebec, Aug. 2002.

[32] 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. of IEEE Int. Conf. on Image Processing*, Barcelona, Sep. 2003.

[33] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proc. of IEEE Int. Conf. on Computer Vision*, Beijing, Oct. 2005.

[34] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. of IEEE Int. Conf. on Computer Vision*, Nice, Oct. 2003.

[35] J. Sivic and A. Zisserman. Video data mining using configurations of viewpoint invariant regions. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Washington DC, Jun. 2004.

[36] A. Smeaton and P. Over. The TREC-2002 video track report. In *Text REtrieval Conference*, Gaithersburg, Nov. 2002.

[37] 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 Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[38] M. Szummer and R.W. Picard. Indoor-outdoor image classification. In *IEEE International Workshop CAIVD, in ICCV'98*, Bombay, Jan. 1998.

[39] A.B. Torralba, K.P. Murphy, W.T. Freeman, and M.A. Rubin. Context-based vision system for place and object recognition. In *Proc. of IEEE Int. Conf. on Computer Vision*, Nice, Oct. 2003.

[40] T. Tuytelaars and L. Van Gool. Content-based image retrieval based on local affinely invariant regions. In *Proc. Visual99*, Amsterdam, Jun. 1999.

[41] 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, 2001.

[42] J. Vogel and B. Schiele. Natural scene retrieval based on a semantic modeling step. In *Proc. of Int. Conf. on Image and Video Retrieval*, Dublin, Jul. 2004.

[43] J. Weston and C. Watkins. Multi-class support vector machines. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, May 1998.

[44] J. Willamowski, D. Arregui, G. Csurka, C.R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *Proc. LAVS Workshop, in ICPR'04*, Cambridge, Aug. 2004.

[45] R. Zhang and Z. Zhang. Hidden semantic concept discovery in region based image retrieval. In *Proc. Conf. on Computer Vision and Pattern Recognition*, Washington, D.C., Jun. 2004.

Pedro Quelhas received the degree in electric engineering in 2001, from the University of Porto, Portugal, and the MRes. in Image Processing and Physics from King's College London in 2002. Currently, he is working towards his PhD. at the IDIAP Research Institute, Switzerland. His research interests are local invariant features, object recognition, and scene classification.

Florent Monay received his M.S. degrees in Microengeneering in 2002 from the Ecole Polytechnique Fédérale in Lausanne (EPFL). Currently, he is a Ph.D. candidate at the IDIAP Research Institute in Martigny, Switzerland. His research interests include multimedia information retrieval, computer vision, and statistical models applied to these domains.

Jean-Marc Odobez (M'03) graduated from the Ecole Nationale Supérieure de Télécommunications de Bretagne (ENSTBr) in 1990, and received his Ph.D degree in Signal Processing from Rennes University, France, in 1994. He then spent one year as a post-doctoral fellow at the GRASP laboratory, University of Pennsylvania, USA. From 1996 until september 2001, he was associate professor at the Université du Maine, France. In 2001, he joined the IDIAP Research Institute as a Senior Researcher, where he is working mainly on the development of statistical models and machine learning algorithms for multimedia and computer vision problems.

Daniel Gatica-Perez (S'01, M'02) received the B.S. degree in Electronic Engineering from the University of Puebla, Mexico in 1993, the M.S. degree in Electrical Engineering from the National University of Mexico in 1996, and the Ph.D. degree in Electrical Engineering from the University of Washington, Seattle, in 2001. He joined the IDIAP Research Institute in 2002, where he is now a senior researcher. His interests include multimedia signal processing and information retrieval, computer vision, and statistical machine learning applied to these domains. He currently is an Associate Editor of the IEEE Transactions on Multimedia.

Tinne Tuytelaars received the MS degree and Ph.D. degree in electrotechnical engineering at the Katholieke Universiteit Leuven in 1996 and 2000 respectively. Currently, she is a post-doctoral researcher of the Fund for Scientific Research Flanders, at the same university. Her research interests are local invariant features, object recognition, wide baseline matching, and scene classification.