Contextual classification of image patches with latent aspect models

Florent Monay[1] monay@idiap.ch
Pedro Quelhas[2] quelhas@fe.up.pt
Jean-Marc Odobez[1, 3] odobez@idiap.ch
Daniel Gatica-Perez[1, 3] gatica@idiap.ch

November 25, 2008

Abstract

We present a novel approach for contextual classification of image patches in complex visual scenes, based on the use of histograms of quantized features and probabilistic aspect models. Our approach uses context in two ways: (1) by using the fact that specific learned aspects correlate with the semantic classes, which resolves some cases of visual polysemy often present in patch-based representations, and (2) by formalizing the notion that scene context is image-specific -what an individual patch represents depends on what the rest of the patches in the same image are-. We demonstrate the validity of our approach on a man-made vs. natural patch classification problem. Experiments on an image collection of complex scenes show that the proposed approach improves region discrimination, producing satisfactory results, and outperforming two non-contextual methods. Furthermore, we also show that co-occurrence and traditional (Markov Random Field) spatial contextual information can be conveniently integrated for further improved patch classification.

1 Introduction

Associating semantic class labels to image regions is a fundamental task in computer vision, useful in itself for image and video indexing and retrieval, and as an intermediate step for higher-level scene analysis [17, 19, 41]. While many image area classification approaches segment an image using all pixels [36] or by predefining a block-based image grid [17, 41], in this work we consider local image patches characterized by viewpoint invariant descriptors [23]. This image representation based on patches, robust with respect to partial occlusion, clutter, and changes in viewpoint and illumination, has shown its applicability in a number of vision tasks [9, 11, 19, 32, 37]. Local invariant regions do not cover the complete image, but they often occupy a considerable part of the scene and divide most of the scene into patches of salient content (Figure 1).

In general, the constituent parts of a scene do not exist in isolation, and the visual context -the spatial dependencies between scene parts- can be used to improve region classification [17,18,21,29]. Two image regions, indistinguishable from each other when analyzed independently, might be discriminated as belonging to the correct class with the help of context knowledge. Broadly speaking, there exists a continuum of contextual models for image region classification. On one end, one would find explicit models like Markov Random Fields, where spatial constraints are defined via local statistical dependencies between class region labels [14, 21], and between observations and labels [17]. The other end would correspond to context-free models, where regions are classified assuming statistical independence between the region labels, and using only local observations [9, 41].

Lying between these two extremes, a type of scene representation of increasing use is the histogram of quantized image patches, referred to as *bag-of-visterms* [31, 33], *bag-of-keypoints* [8], *bag-of-features* [24], or *bag-of-codewords* [11, 13] in the literature. This representation is obtained by sampling local regions

[1] Idiap Research Institute, Martigny, Switzerland
[2] INEB-Instituto de Engenharia Biomedi, Campus da FEUP, Porto, Portugal
[3] Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

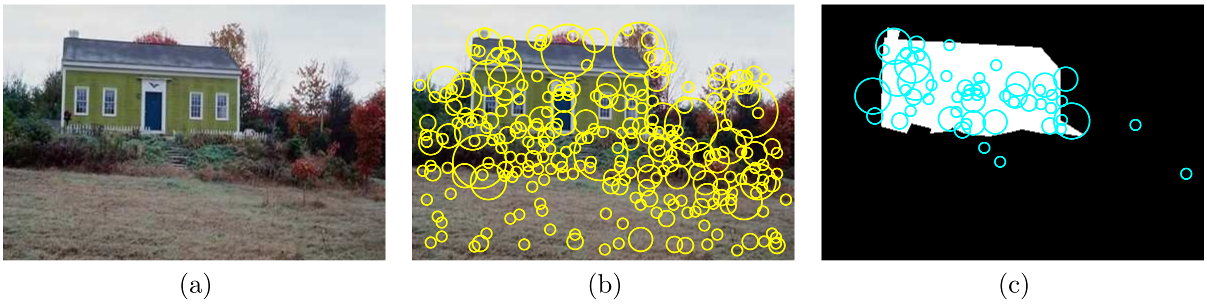


Figure 1: (a) a visual scene, (b) scene patches: local invariant regions in yellow, (c) patches are classified with our method either as man-made (in blue) or nature (not shown), and superimposed on a manual image area classification (in white).

in an image and quantizing them into a finite set of *patches* according to their visual appearance, storing the patch occurrence in the image in the form of an histogram. On one hand, unlike explicit contextual models, spatial neighboring relations in this representation are discarded, and any ordering between the image regions disappears. On the other hand, unlike point-wise models, although the image regions are still local, the scene is represented collectively. This can explain why, despite the loss of strong spatial contextual information, this type of representation has been successfully used in a number of problems, including object matching [38], object categorization [37, 43], scene classification [4, 11, 32], and scene retrieval [41].

As a collection of discrete data, the histogram of patches is suitable for probabilistic models that capture a different form of context which is implicitly captured through patch co-occurrence. These models, originally designed for text collections (documents composed of terms), use discrete hidden *aspect* variables to model the co-occurrence of terms within and across documents. Examples include Probabilistic Latent Semantic Analysis (PLSA) [15] and Latent Dirichlet Allocation (LDA) [3]. We have recently shown that the combination of PLSA and histogram of quantized invariant local descriptors can be successfully used for global scene classification [31,32]. Given an unlabeled image set, PLSA captures aspects that represent the class structure of the collection, and provides a low-dimensional representation useful for classification. Similar conclusions with an LDA related model were reached in [11].

In this paper, we address the problem of classifying image regions into semantic classes (see Figure 1) based on their associated patch number[1]. The main challenge for this task is that patches are not class-specific. As shown in Figure 2, image regions quantized into the same patch can appear in both man-made and nature views. This situation, although expected since quantized patch construction does not make use of class label information, constitutes a problematic form of visual polysemy. In this paper, we propose to take advantage of the context in which each patch appears, characterized by the patch histogram itself, to improve the classification of the corresponding image regions. Our contributions can be summarized as follows:

1. We show that the above mentioned aspect models can be directly applied to patch classification, since specific aspects, although learned without class information, correlate with the classes of interest. These aspects can be easily labeled by hand or using a labeled image dataset, and used to classify their most likely patches accordingly.

2. The interpretation of a particular patch depends on what the other patches in the same image are, and this co-occurrence context is precisely captured by the estimated aspect mixture weights. We propose to formally include this contextual information in a new aspect model, so that even though patches appear in multiple classes, the information about the other patches in the same image can be used to improve discrimination (Figure 2).

3. We present results on a *man-made* vs. *natural* image regions classification task, and show that the contextual information learned from co-occurrence improves the performance compared to a

[1] Throughout the paper, the term patch will mainly be used to denote an image region, and sometimes to denote the discrete index obtained from quantizing a local image descriptor of the patch. In case of ambiguity, we will use the term quantized patch or patch number to denote the later.

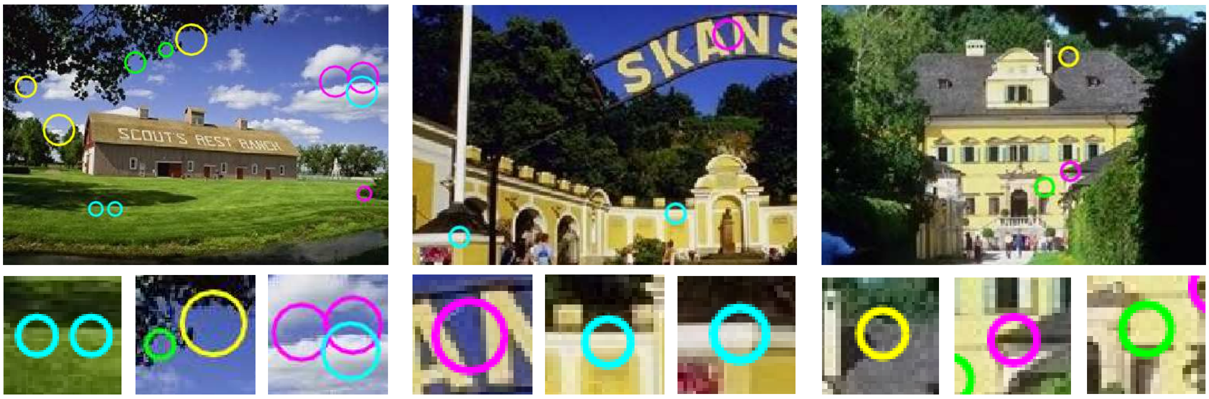


Figure 2: Image local regions can have different scene class labels depending on the image in which they are found. Left: various patches (4 different colors, same color means same patch number) that occur on *natural* parts of an image. Center and Right: the same patches occur in man-made structures. All these regions are correctly classified by our approach, switching the class label for the same patch depending on the context.

non-contextual approach. In our view, the proposed approach constitutes an interesting way to model visual context that could be applicable to other problems in computer vision.

4. We show, through the use of a Markov Random Field model, that standard spatial context can be integrated, resulting in an improvement of the final classification of image regions.

The paper is organized as follows. Section 2 reviews the closest related work. Section 3 presents our approach to local image patch classification. Section 4 introduces the image representation. Section 5 introduces the concept of an image as a mixture of latent aspects that is extended in Section 6 for contextual local patch classification. Section 7 discusses the two baseline models. Section 9 reports our results. Section 10 concludes the paper.

2 Related work

Image region classification is a research field that has been developed for many years. Generally speaking, there are two main approach directions to the problem: classic pixel based image segmentation and image region classification.

Classic image segmentation is defined as a process of partitioning the image into non-intersecting regions, such that each region is homogeneous and no union of two adjacent regions is homogeneous [30]. The main issue is defining the property by which homogeneity is imposed. In most cases, the properties on which segmentation is based are gray-scale, color, texture, or a combination of those properties. Image segmentation defined this way is performed on each image independently. A review of traditional segmentation approaches is given in [30]. Many more alternatives have been proposed. For instance, Carson et al. [7] present a blob-based segmentation method that models the color, texture and position of all the pixels in a given image with a Gaussian mixture model (GMM), and attribute the label of its most likely GMM component to each pixel. This creates roughly homogeneous image regions called blobs, that are used for image retrieval, allowing the user to query the database at the blob level instead of the image level.

We consider the perspective on image region classification which is based on automatically defined patches. As we will show this allows the regional classification of images based on class labels that are predefined and applicable to the whole database, and not based on an homogeneity criterion of the regions in an image. The region descriptors are classified into categories, and the density of the region class labels gives a regional classification of the image. We present a selection of image regional classification models that are based on class labels in the next paragraphs, with regions that cover the whole image [10, 17, 40–42], or only a part of it [9, 19, 37].

The work in [10] relies on the Normalized Cuts segmentation algorithm [35] to segment the image into regions that are then quantized. Derived from the machine translation literature, an Expectation-Maximization (EM) estimates the probability distributions linking a set of words and blobs. Once the model parameters are learned, words are attached to each region. This *region naming* process is comparable to image segmentation.

Extending the Markov Random Field (MRF) model, Kumar and Herbert proposed a Discriminative Random Field (DRF) model that includes neighborhood interactions in the class labels, as well as at the observation level. They apply the DRF model to the segmentation of man-made structures in natural scenes [17], with an extraction of images features based on a grid of blocks that fully covers the image. The DRF model is trained on a set of manually segmented images, and then used to infer the segmentation into the two target classes.

Using a similar grid layout, Vogel and Schiele presented a two-stage classification framework to perform scene retrieval [41] and scene classification [42]. This work performs an implicit scene segmentation as an intermediate step, classifying each image block into a set of semantic classes such as *grass*, *rocks*, or *foliage*.

To include global shape prior information in an MRF-based model formulation, Kumar et al. proposed an MRF part-based segmentation model, referred to as *ObjCut*, which represents object by means of segmented parts [16]. This requires the explicit encoding of the spatial information relating parts and also the modeling of their deformations. The use of regions in this case reduces the invariance to occlusion, and the modeling has a high computational cost. Furthermore, the object to model must be composed of discriminative parts with known spatial relationships, which is not the case for scenes.

In [9], invariant local descriptors are used for an object detection task. All region descriptors in the training set are modeled with a Gaussian Mixture Model (GMM). A subset of the mixture components is then selected based on their estimated class likelihood ratio or mutual information, that are then used to classify new regions based on their local descriptors. In this non-contextual approach, new descriptors are independently classified into object or background regions, without taking the other descriptors in the same image into consideration. A similar approach introducing spatial contextual information through neighborhood statistics of the GMM components collected on training images is proposed in [19], where the learned prior statistics are used for relaxation of the original region classification.

Leibe et al. proposed an implicit object model based on local invariant descriptors that jointly learns the discriminant descriptors for an object and their spatial relationships [20]. Once again, this approach implies an existing spatial layout of the object parts which does not exist in the case of scenes.

As an extension to local descriptors' representation of images, probabilistic aspect models have been recently proposed to capture descriptors co-occurrence information with the use of a hidden variable (latent aspect). The work in [11] proposed a hierarchical Bayesian model that extended LDA for global categorization of natural scenes. This work showed that important patches for a class in an image can be found. However, the problem of local image patch classification was not addressed. The combination of local descriptors and PLSA for local patch classification has been illustrated in [37]. However this work has two limitations. First, patches were classified into aspects, not classes, unless we assume as in [37] that there is a direct correspondence between aspects and semantic classes. This seems however a over-simplistic assumption in general. Secondly, evaluation was limited, e.g. [37] does not conduct any objective performance evaluation.

To model both the object and the scene in an image, Russel et al. [34] proposed to use regions resulting from multiple unsupervised image segmentations to represent an image as an aggregate of sub-images. These sub-images are represented with bag-of-visterms and modeled with an latent aspect model. Starting from multiple image segmentations to maximize the chance that some segmented regions will correspond to actual objects is an interesting approach. There is however no guarantee that this will be true in general, and we therefore model images at the scale of patches in our work to ensure that no initial segmentation step will harm the image representation.

A preliminary version of our work first appeared in [28]. Inspired by our work, Verbeek et al. proposed the extension of aspect modeling by integrating spatial models [40]. The proposed approach introduces spatial coherence to the aspect model improving segmentation. However, the training of the latent aspect becomes limited to using labeled data, losing the possibility of learning visual co-occurrence from unlabeled data.

Unlike previous approaches, we propose a formal way to integrate the latent aspect modeling, learned in an unsupervised way from unlabeled data, in the class information, and conduct a proper performance

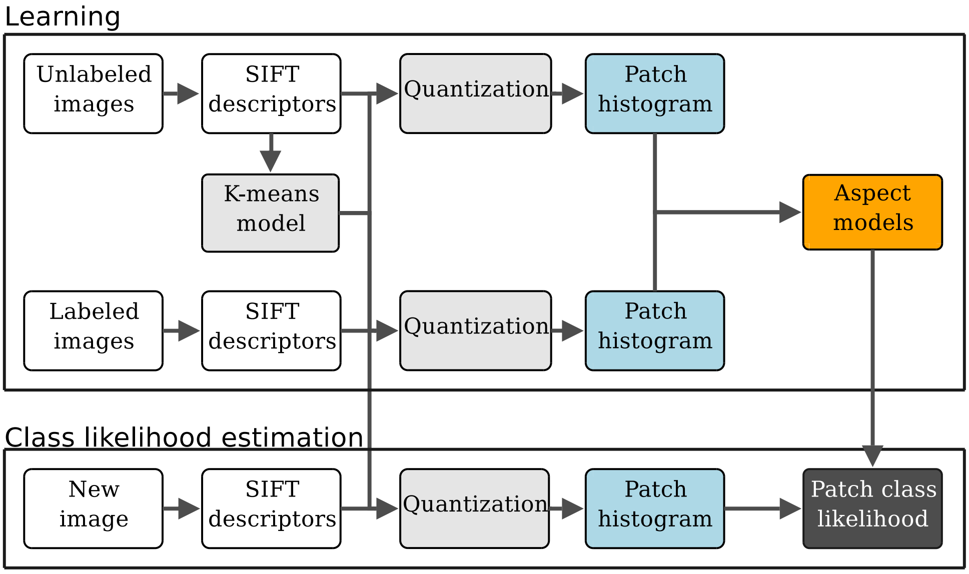


Figure 3: Our aspect models rely on a patch-based image representation, obtained by a K-means quantization of SIFT image region descriptors. The class likelihood of patches extracted from a new image is estimated from the previously seen labeled images.

evaluation, validating our work with a comparison to a state-of-the-art baseline method. In addition, we explore the integration of the more traditional spatial MRF model into our system and compare the obtained results.

In the final stage of preparation of this manuscript, new models were put forward to segment images by combining latent aspect models with quantized local patches. Fei-fei et al. presented a latent aspect model that assumes that each region of an image, obtained with an unsupervised segmentation algorithm in a first step, are generated from a single aspect [6]. Regions are not modeled as separate documents, but as building parts of a given image that is itself defined by a mixture of aspects, contrarily to [34]. Liu and Chen proposed to explicitly combine a latent aspect model with a known supervised segmentation algorithm [22]. The segmentation algorithm and the aspect models are linked through a new variable that distinguishes foreground from background patches. This variable is successively obtained from the segmentation algorithm and then considered as an observed variable in the aspect model. A new segmentation is obtained when the aspect model is learned and this process iterates until the final segmentation is obtained.

3 Scene patch classification

The aspect models that we present in this paper allow to classify image regions into two classes, based on an estimated patch class likelihood taking advantage of the availability of a patch histogram. The method can be applied to image collection of regions defined randomly, by a regular grid (with or without overlap), or obtained with an interest point/region detector. Depending on what the considered image regions are, the resulting spatial distribution of class labels can produce local image classification with no label overlap (e.g. when using grid patches) [17, 41, 42], or a density-based image patch classification (when using interest point detectors) [9, 19]. In the later case, as shown on Figure 1, the classification of patches obtained by an interest point detector produces a sparse regional image classification. However, one advantage of using an interest point detector is the identification of stable regions may exhibit better correspondence across the images than an arbitrary grid image division. In this paper, we decided to rely on an interest point detector to sample specific types of image regions to be classified, but the technique can be applied to any other form of region selection scheme.

As shown on Figure 3, our approach relies on the quantization of local region descriptors into a fixed number of patches using the K-means clustering algorithm. Compared to [9] and [19], this quantization step simplifies the image representation from an undefined number of region descriptors per image to an histogram of patch labels. In addition, it allows to define a patch co-occurrence context of an image as a simple histogram, which can be further analyzed with an aspect model formulation. The patch histogram representation is discussed in details in Section 4.

Classification principle: likelihood ratio.

We rely on likelihood ratio computation to classify each patch v of a given image d into a class c. The ratio is defined by

$$LR(v) = \frac{P(v|c = \text{man-made})}{P(v|c = \text{natural})}, \tag{1}$$

where the probabilities will be estimated using different models of the data, as described in Section 6, and the classification rule is :

$$LR(v) > T \Rightarrow v \in \text{man-made}, \tag{2}$$

where T is a threshold value. Thus all image regions associated to the same patch will be classified in the same category according to the rule 2. Note that alternatively, we could have considered as classification rule a ratio based on $P(c|v)$. The only difference with respect to using $LR(v)$ is to multiply the threshold value T by the constant $P(c = man - made)/P(c = natural)$.

4 Image representation

In the following, we describe and further justify the four steps that we take to build our image representation: (i) detection of interest points/patches, (ii) computation of local descriptors, (iii) local descriptor quantization, and (iv) construction of the patch histogram.

4.1 Detection of interest points

The goal of the interest point detector is to automatically extract characteristic points from a given image, which are invariant to some geometric and photometric transformations. These points define image regions which are also invariant to the same transformations. Invariance is an important property since it ensures that given an image and its transformed version, equivalent image patches will be extracted from both and the resulting image representation will be the same (within a certain estimation error).

Different point detectors have been proposed to extract regions of interest in images [23, 39]. 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. However, the increase in invariance also means that to different points can become more similar after invariance regularization. In this way, we must also restrain invariance since a big increase in the degree of invariance may remove information about the local image content that is valuable for classification.

In this work, we use the difference of Gaussians (DOG) point detector [23]. 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 was shown to perform well in comparison studies previously published [25, 26], and also since we found it to be a good choice in practice for the task at hand, performing competitively compared to other detectors [32]. The DOG detector is also faster than similarly performing, fully affine-invariant ones [39],

4.2 Computation of local descriptors

Local descriptors are computed over the image region defined by each interest point that is automatically identified by the local interest point detector. These descriptors characterize the image content of each region in a compact way. In this work we use the SIFT (Scale Invariant Feature Transform) feature as local descriptors [23]. This choice was motivated by several publications [11, 25], where SIFT was found to work best. This descriptor is based on the gray-scale gradient information of images, and was shown to perform best in terms of specificity of region representation and robustness to image transformations [25]. SIFT features are local histograms of edge directions computed over different parts of the region of interest, capturing the structure of the local image patch. In [23], it was shown that the use of 8 orientation directions and a grid of 4x4 parts gives a good compromise between descriptor size and accuracy of representation (see Figure 4), what gives a feature vector of size 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 in the elaboration of the SIFT feature vector are then done with respect to this dominant orientation.

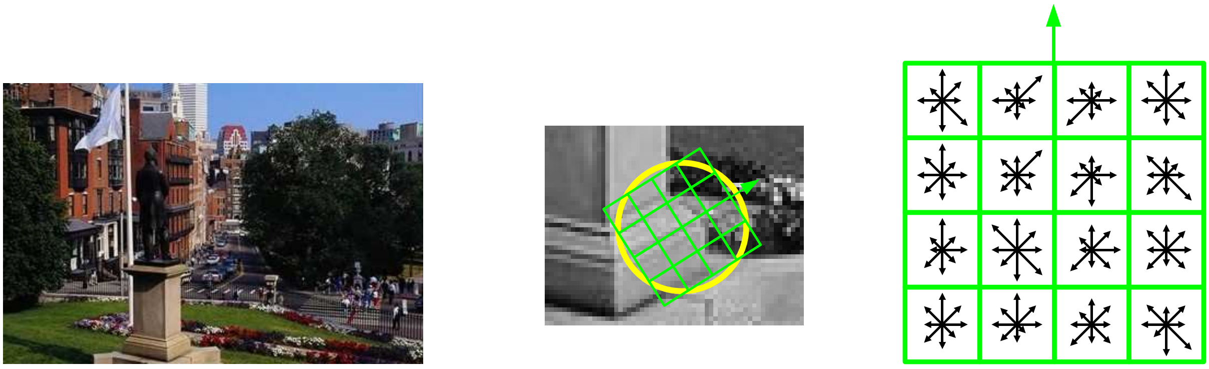


Figure 4: The Scale Invariant Feature Transform (SIFT) descriptor. The detected regions are segmented into a 4×4 grid, and each square is represented by an eight-bin histogram of the edge directions in this region, resulting in a description vector of dimension 128.

4.3 Local descriptor quantization

After the interest point detection and the computation of descriptors, an image is represented as a set of SIFT features characterizing the gray-scale texture of its regions of interest. We propose to quantify the descriptors to obtain a fixed size, compact representation of the image. A vocabulary of quantized descriptors $\mathcal{V}$- referred to as *patches* in this paper - is constructed by learning a K-means model from a set of local descriptors extracted from the training images, keeping the estimated $N_{\mathcal{V}}$ means as patches. New local descriptors s are mapped to the closest patch v in the vocabulary $\mathcal{V}$ according to the nearest neighbor rule:

$$s \longmapsto Q(s) = v_i \iff \text{dist}(s, v_i) \leq \text{dist}(s, v_j) \qquad \forall j \in \{1, \ldots, N_{\mathcal{V}}\} \tag{3}$$

where $N_{\mathcal{V}}$ denotes the size of the patch set. We used the Euclidean distance in the clustering (and in Equation 3) and choose the number of clusters depending on the desired vocabulary size. The choice of the Euclidean distance to compare SIFT features is common [23].

Technically, the quantization of similar local descriptors into a single patch 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 descriptors into patches.

Furthermore, local descriptors will be considered as distinct whenever they are mapped to different patches, 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, and that any distance defined over their representation (e.g. strings in the case of text) carries no semantic meaning.

Figure 5 shows some examples of clusters of the SIFT descriptors. All of the examples of each cluster get the same label, and so get represented by the same patch. The patch number 157 represents a step function that might not be very specific to any of the *man-made* or *natural* image regions. On the contrary, the patches 240 and 14 represent cornered/squared structures that should mostly occur in *man-made* structures. Similarly, the samples from patch 661 contain high frequencies that seem most likely to occur in *natural* structures.

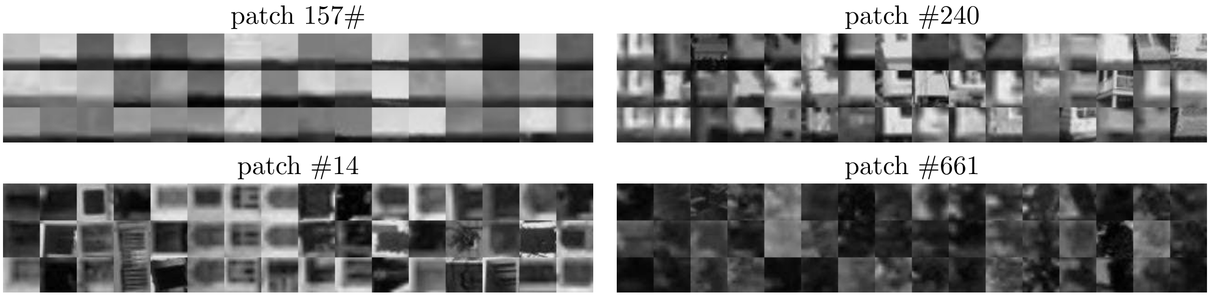


Figure 5: Four examples of randomly selected image regions clustered into the same patch number, out of 1000 obtained by the K-means quantization.

4.4 Patch histogram

After the feature quantization step, the image is reduced as a set of patches taken from a fixed size patch vocabulary, that can be encoded as a patch histogram according to:

$$h(d) = (h_i(d))_{i=1..N_{\mathcal{V}}}, \text{ with } h_i(d) = \mathrm{n}(d, v_i) \tag{4}$$

where $\mathrm{n}(d, v_i)$ denotes the number of occurrences of patch v_i in image d. The construction of the patch histogram is illustrated in Figure 6. The patch histogram contains no information about spatial relationship between patches, similarly to the bag-of-words text representation: even though word ordering contains a significant amount of information about the original data, it is completely removed from the final document representation.

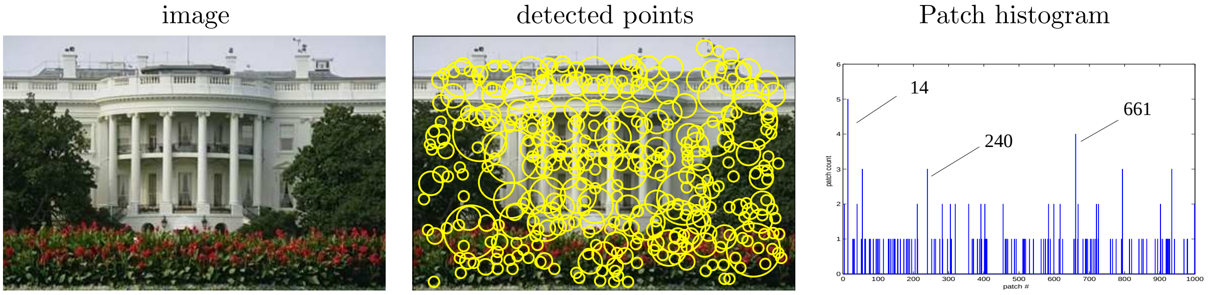


Figure 6: Construction of the patch histogram representation. Image regions are detected with the Difference of Gaussians (DoG) detector, their SIFT representation are extracted and then quantized to build the patch histogram.

5 Scenes as mixtures of aspects

The concept of aspect models for images has been recently applied to scene [4,32,33] and object [12,27] categorization tasks, using the estimated distribution over aspects as a feature extraction process, or directly as a classifier. Under the assumption of an aspect model, an image can be seen as a mixture of unobserved (latent) aspects, that are defined by consistent co-occurrences of image patches (or their features) within the image collection. A latent aspect z_k is thus represented by its conditional distribution over patches $P(v \mid z_k)$, and an image d_i is represented by the conditional distribution over aspects $P(z \mid d_i)$.

5.1 Scene modeling with PLSA

Several latent aspect models such as Probabilistic Latent Semantic Analysis (PLSA) [15], Latent Dirichlet Allocation (LSA) [3], and Multinomial PCA (MPCA) [5] have been proposed in the literature for discrete

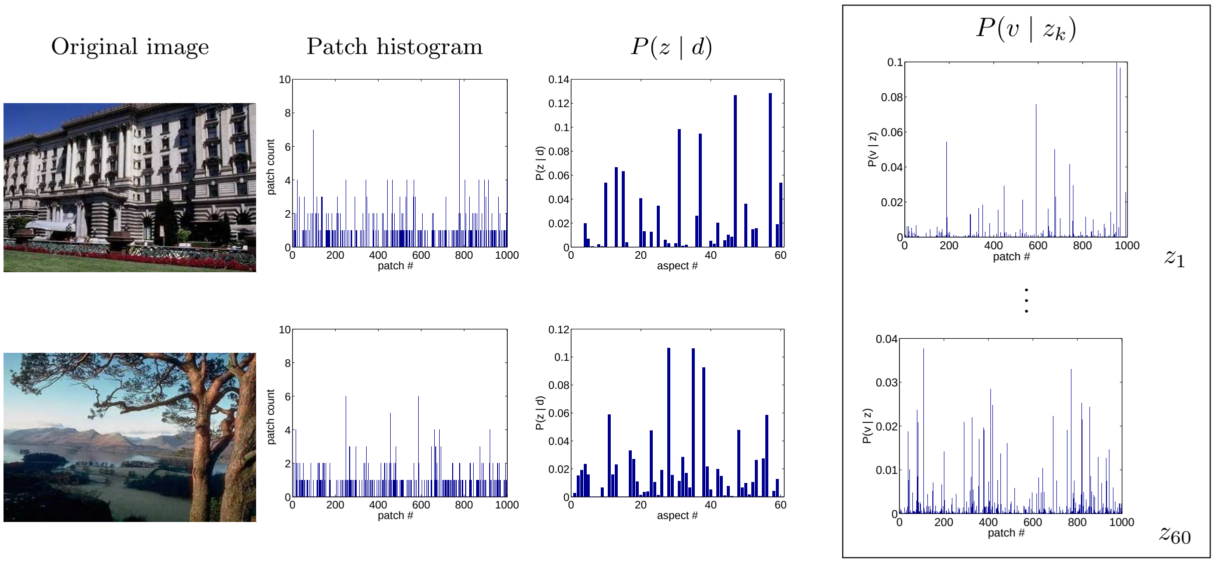


Figure 7: Two images and their decomposition into a mixture of $N_A = 60$ aspects, estimated by the PLSA model. The second column is the histogram of 1000 patches corresponding to the image on the same row, the third column shows the estimated distribution over aspects given the patch histogram. The right column represents the N_A conditional distributions over patches given the aspects z_k.

components analysis. In this work, we consider the PLSA model [15], which assumes each occurrence of the patch v_j to be independent from the image it belongs to given the latent variable z_k, and corresponds to the joint probability expressed by:

$$P(v_j, z_k, d_i) = P(d_i)P(z_k \mid d_i)P(v_j \mid z_k). \tag{5}$$

The joint probability of the observed variables is the marginalization over the N_A latent aspects z_k as expressed by:

$$P(v_j, d_i) = P(d_i)\sum_{k=1}^{N_A} P(z_k \mid d_i)P(v_j \mid z_k). \tag{6}$$

The multinomial distributions $P(z \mid d_i)$ and $P(v \mid z_k)$ are estimated with an EM algorithm on a set of training documents. As an illustration, the Figure 7 shows the distribution over aspects for two images, for an aspect model trained on a collection of 6600 images of landscape and city images. The conditional distributions of patches given the $N_A = 60$ aspects are represented on the right column of Figure 7, representing an aspect by its specific patch co-occurrence pattern. We see in Figure 7 that the patch histogram representations of the two images are modeled by two dissimilar distributions over aspects, reflecting their differences in content. The two images are composed of different patch co-occurrences that exist in the image collection, resulting in different image-dependent contexts.

The aspect indices have no intrinsic relevance to a specific class, given the unsupervised nature of the PLSA model learning. We can however inspect each aspect to observe the meaning that they may have in terms of our target classes. Aspects can be conveniently illustrated by their most probable images in a dataset. Given an aspect z, images can be ranked according to:

$$P(d|z) = \frac{P(z|d)P(d)}{P(z)} \propto P(z \mid d), \tag{7}$$

where $P(d)$ is considered as uniform. Figure 8 displays the 10 best-ranked images for a given aspect to illustrate its potential 'semantic meaning'. The top-ranked images representing aspect 55 and 22 all clearly belong to the *natural* class, while the top-ranked images for aspect 50, 10, and 37 contain a large majority of *man-made* structures. Aspect 12 seems to be mainly related to horizon/panoramic scenes, and contains landscape images only (top 10 images). However, as aspects are identified by analyzing

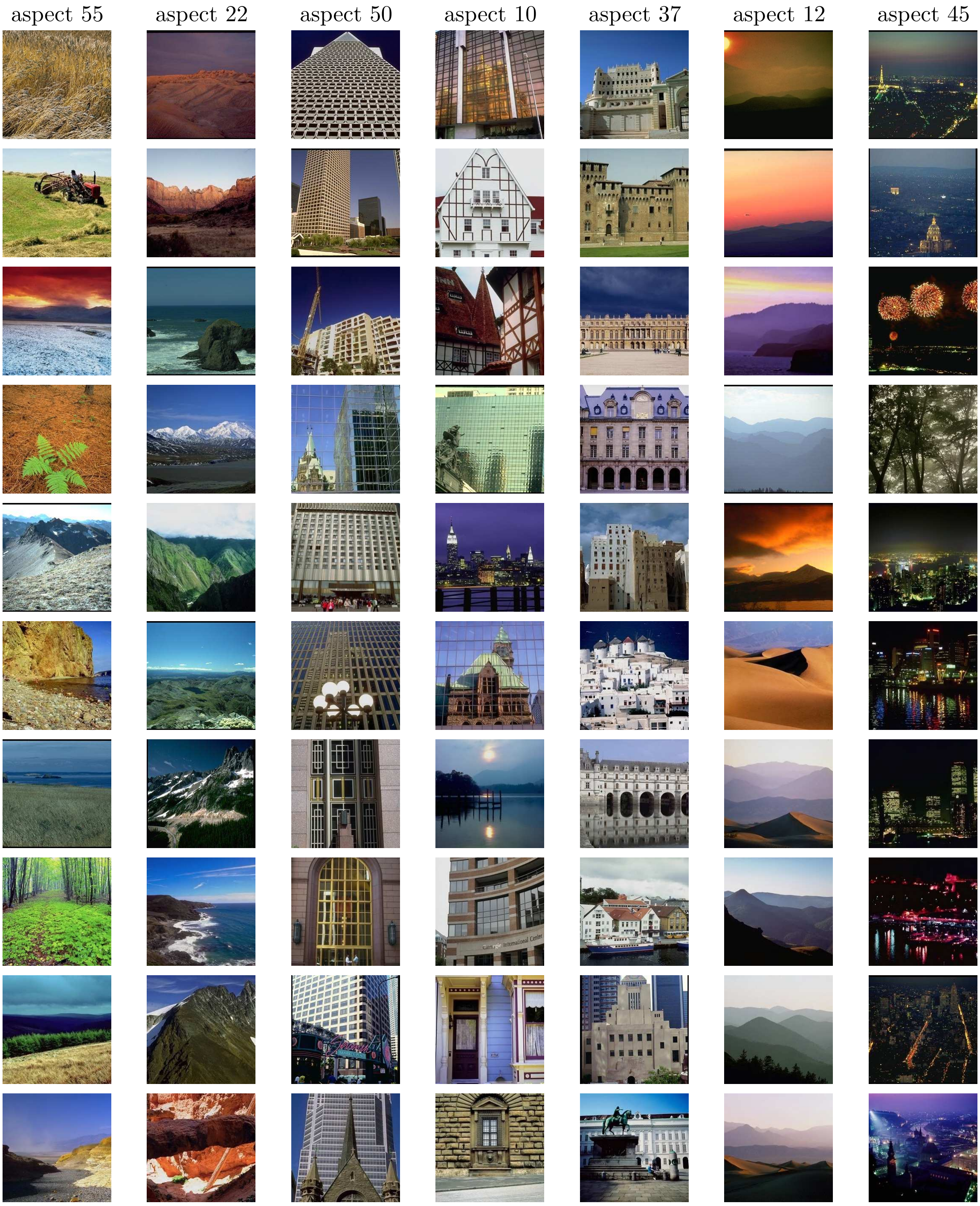


Figure 8: Illustration of seven aspects out of 60 learned by the PLSA model on a set of 6600 landscape and city images. The 10 top-ranked images for each aspects are displayed, showing a correspondence between the aspects and the *man-made* (aspects 50, 10 and 37) and *natural* (aspects 55, 22 and 12) classes.

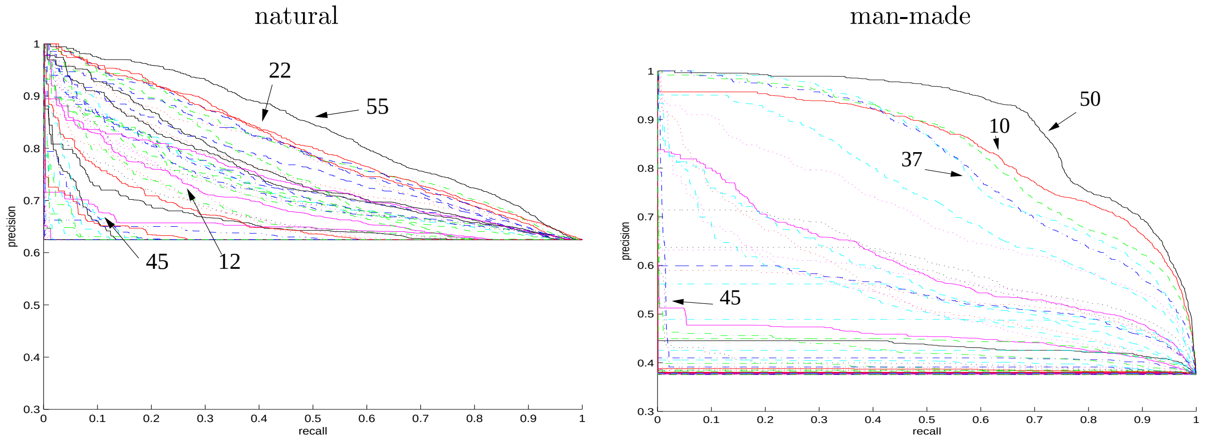


Figure 9: Precision/recall curves for the image ranking based on each of the 60 individual aspects, relative to the *natural* (left) and *man-made* (right) query. Each curve represents a different aspect. Floor precision values correspond to the proportion of *natural* (resp. *man-made*) images in the dataset.

the co-occurrence of visual patterns within local patches, they may be consistent from this point of view without allowing for a direct semantic interpretation as shown on Figure 8 for the aspect 45.

To further confirm the connection between the learned aspects and the target classes we can measure objectively their relationship by defining the *Precision* and *Recall* paired values w.r.t a given label at rank r 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. Note here that for this experiment, we assume that images are only associated with one class label, although they may contain some content (and patches) belonging to the other class. The precision/recall curves associated with each aspect-based image ranking considering either the *natural* or the *man-made* queries are shown in Figure 9. Those curves prove that some aspects are clearly related to the two classes, and confirm the observations made previously with respect to the aspect correspondences. As expected, aspect 45 does not appear in either the *man-made* or the *natural* top precision/recall curves. The *natural* related ranking of aspect 12 does not hold as clearly for higher recall values, because the pattern of patch co-occurrences appearing in horizons that it captures is not exclusive to the *natural* class.

5.2 Mapping aspects to local image patches

As we have shown, images can be modeled as mixtures of aspects, and some aspects correlate with the *man-made* or the *natural* classes. The conditional distribution of patches given an aspect $P(v|z)$ could be exploited for the classification of image regions in an image (given their patch label), as far as as a class label is attached to the aspects. Based on the learned conditional distributions of patches given aspects, the most likely aspect can be attributed to a given patch according to:

$$\begin{aligned} z_{v_j} &= \underset{z}{\operatorname{argmax}}\ (P(z|v_j)) \\ &= \underset{z}{\operatorname{argmax}}\ (\frac{P(v_j|z)P(z)}{P(v_j)}) = \underset{z}{\operatorname{argmax}}\ P(v_j|z), \end{aligned} \tag{8}$$

where we have assumed that the distribution over the latent aspects $P(z)$ is uniform. In Figure 10, we show two examples of image region classification based on the concept of mixture of aspects. Based on the average precision (AP) measure of the ranking illustrated in Figure 9, we first select the ten aspects that are the more closely related to the *man-made* class and the ten aspects that are the more closely related to the *natural* class. Restricting the aspect attribution to these 20 *man-made* and *natural* aspects, each patch can be independently classified as a *man-made* or a *natural* descriptor based on

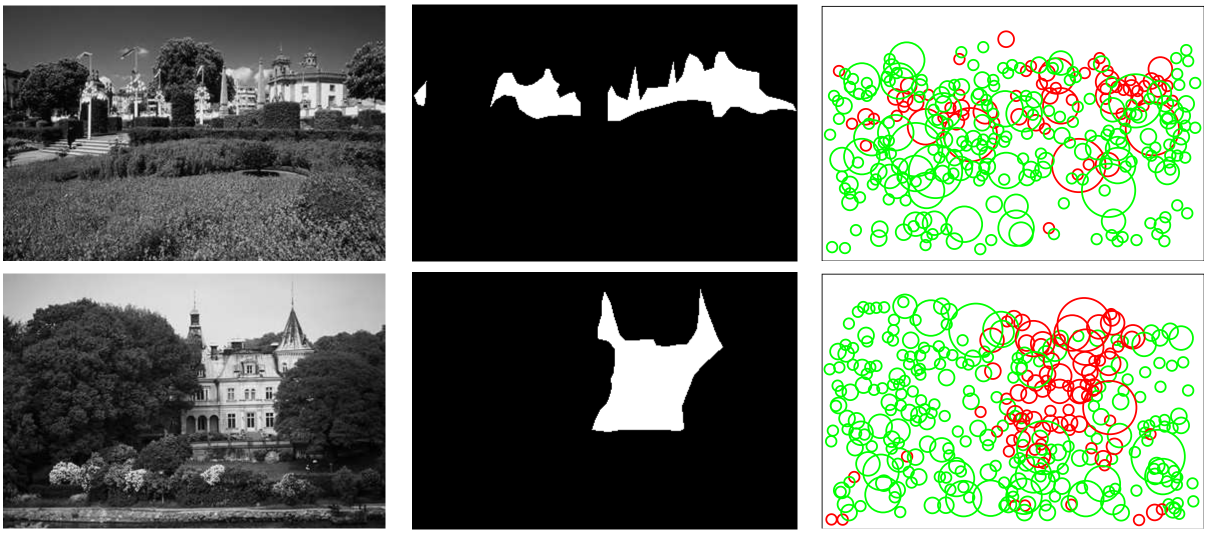


Figure 10: Classification of local image patches based on the 10 aspects that are the more closely related to the *man-made* class, and the 10 aspects that are the more closely related to the *natural* class. The first column is the original image, the second column is the ground-truth image area classification (white is *man-made*, black is *natural*), and the last column is result of the patch classification. Red circles correspond to patches classified as *man-made*, green circles correspond to patches classified as *natural* (see text). The respective densities of red and green points show a good correspondence with the ground-truth image area classification.

Equation 8. These two examples show a reasonable match between the ground-truth patch classification and the density of red and green points. The unsupervised learning based on co-occurrence thus allows to identify *man-made* and *natural* latent aspects in the data, that can be later used to classify patches (and their corresponding image regions) into these two categories.

Based on this idea, we present two aspect models that extend the Probabilistic Latent Semantic Analysis (PLSA) model [15] for image patch classification in the next Section.

6 Aspect models for patch classification

As introduced in Section 3, our goal is to classify image regions based on the estimated class-likelihood ratio of their corresponding patches, as described in Equation 1. In the following, we propose two aspect models that estimate patch class-likelihoods based on the decomposition of scenes in a mixture of aspects. The observed data is composed of patch, document and class triplets (v, d, c) for each patch occurrence in a labeled training set.

The first aspect model classifies patches independently of the image they belong to, and can be thus seen as a probabilistic formulation of the idea presented at the end of Section 5, where the assumption was that an aspect could only be associated with one class (i.e. $P(z|c) = 0$ or 1). The second model takes full advantage of the patch histogram context, and allows to estimate patch class-likelihoods that depends on the image that is considered.

6.1 Aspect model 1

The first model associates a hidden variable $z \in \mathcal{Z} = \{z_1, \ldots z_{N_A}\}$ with each observation leading to the joint probability defined by

$$\begin{aligned} P(c,d,z,v) &= P(v|z,d,c)P(z|d,c)P(d|c)P(c) & (9) \\ &= P(v|z)P(z|d)P(d|c)P(c). & (10) \end{aligned}$$

This model introduces two conditional independence assumptions. The first one, traditionally encountered in aspects models, is that the occurrence of a patch v is independent of the image d it belongs to, given an aspect z. The second assumption is that the occurrence of aspects is independent of the class the patch belongs to, i.e. $P(z|d,c) = P(z|d)$. Note that in the above equation, the class label refers to the class of one patch. Thus, different class labels can be associated with a given document, and the term $P(d|c)$ reflects the degree to which an image indirectly belongs to a given class given its patches. The parameters of this model are learned using the maximum likelihood (ML) principle [15]. The optimization is conducted using the Expectation-Maximization (EM) algorithm, allowing us to learn the aspect distributions $P(v|z)$ and the mixture parameters $P(z|d)$.

Notice that, given our model, the EM equations do not depend on the patch class label. Besides, the estimation of the class-conditional probabilities $P(d|c)$ does not require the use of the EM algorithm. We will exploit these points to train the aspect models on a large dataset (denoted $\mathcal{D}$) where only a small part has been manually labeled at the image level (we denote this subset by $\mathcal{D}_{lab}$). This labeling at the image level allows to quickly annotate a large number of patches as man-made or natural , but does not implies that images have one class in general. We assume that patches have a class label.

Regarding the class-conditional probabilities, as the labeled set is only composed of man-made-only or natural-only images, we simply estimate them according to:

$$P(d|c) = \begin{cases} 1/N_c & \text{if } d \text{ belongs to class } c \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where N_c denotes the number of images belonging to class c in the labeled set $\mathcal{D}_{lab}$. Given this model, the likelihood we are looking for (cf. Equation 1) can be expressed as

$$P(v|c) = \sum_{l=1}^{N_A} P(v, z_l|c) = \sum_{l=1}^{N_A} P(v|z_l)P(z_l|c), \tag{12}$$

where the conditional probabilities $P(z_l|c)$ can in turn be estimated through marginalization over labeled documents,

$$P(z_l|c) = \sum_{d\in\mathcal{D}_{lab}} P(z_l, d|c) = \sum_{d\in\mathcal{D}_{lab}} P(z_l|d)P(d|c). \tag{13}$$

These equations allow us to estimate the likelihood ratio as defined by Equation 1. Note that this model extends PLSA by introducing the class variable [15].

6.2 Aspect model 2

From Equation 12, we see that, despite the fact that the above model captures co-occurrence of the patches in the distributions $P(v|z)$, the context provided by the specific image d has no direct impact on the likelihood. To explicitly introduce this context knowledge, we propose to evaluate the likelihood ratio of patches conditioned on the observed image d,

$$LR(v,d) = \frac{P(v|d, c = \text{man-made})}{P(v|d, c = \text{natural})}. \tag{14}$$

The evaluation of $P(v|d,c)$ can be obtained by marginalizing over the aspects,

$$P(v|d,c) = \sum_{l=1}^{N_A} P(v, z_l|d,c) = \sum_{l=1}^{N_A} P(v|z_l)P(z_l|d,c), \tag{15}$$

where we have exploited the conditional independence of patch occurrence given the aspect variable. Under model 1 assumptions, $P(z_l|d,c)$ reduces to $P(z_l|d)$, which clearly shows the limitation of this model to introduce both context and class information for patch classification. To overcome this, we assume that the aspects depend on the class label as well. The parameters of this model are the aspect multinomial $P(v|z)$ and the mixture multinomial $P(z|d,c)$, which could be estimated from labeled data by EM as before. However, as our model is not fully generative [3], only $P(v|z)$ can be kept fixed, and we would have to estimate $P(z|d_{new}, c)$ for each new image d_{new}. We propose to separate the contributions to the aspect likelihood due to the class-aspect dependencies, from the contributions due to the image document-aspect dependencies. Thus, we propose to approximate $P(z_l|d,c)$ as

$$P(z_l|d,c) \propto P(z_l|d)P(z_l|c), \tag{16}$$

where $P(z_l|c)$ is still obtained using Equation 13. The complete expression is given by

$$P(v|d,c) \propto \sum_{l=1}^{N_A} P(v|z_l)P(z_l|c)P(z_l|d). \tag{17}$$

The main difference with Equation 12 is the introduction of the contextual term $P(z_l|d)$, which means that patches will not only be classified based on them being associated to class-likely aspects, but also on the specific occurrence of these aspects in the given image.

Inference on new images

With aspect model 1 (and also with empirical distribution, cf. baseline model in Section 7), the patch classification decision is taken once for all at training time, through the patch co-occurrence analysis on the training images. Thus, for a new image d_{new}, the extracted patches are directly assigned to their corresponding most likely class label. For aspect model 2, however, the likelihood-ratio $LR(v, d_{new})$ (Equation 14) involves the image dependent aspect parameters $P(z|d_{new})$ (Equation 17). Given our approximation (Equation 16), these parameters have to be inferred for each new image, in a similar fashion as for PLSA [15]. $P(z_l|d_{new})$ is estimated by maximizing the likelihood of the patch histogram of d_{new}, fixing the learned $P(v|z_l)$ parameters in the Maximization step.

7 Baseline Models

We propose two complementary baseline models. The first baseline directly uses the empirical patch class-conditional distribution to classify new patches, the second learns a model from the region descriptors themselves, without quantification.

7.1 Empirical class-conditional patch distribution

Given a set of training data, the ratio in Equation 1 can simply be estimated using the empirical distribution of patches, as done in [9]. More precisely, given a set of manually segmented images $\mathcal{D}$into man-made and natural regions (e.g. Figure 1 (c)), $P(v|c)$ is estimated as the number of times the patch v appears in regions of class c, divided by the total number of visterms of class c in the training set. Note that the class conditional probabilities $P(c|v)$ could have been considered instead. This would have modified the estimated likelihood threshold value T_{EER} by $P(c = man{-}made)/(1{-}P(c = man{-}made))$. The class conditional probabilities $P(c|v)$ are shown in Figure 11, indicating that there is a substantial amount of polysemy. Patches can simultaneously have a high probability given both classes (e.g. for instance note that all patches appear at least 15% in the *natural* class).

Empirical estimation of probabilities is simple but may suffer from several drawbacks. A first one is that a significantly large amount of labeled training data might be necessary to avoid noisy estimates, especially when using large vocabulary sizes. A second one is that such estimation only reflects the individual patch occurrences, and does not account for any kind of relationship between them. Patches however correspond to regions extracted from full images, and therefore should be better interpreted in this context. In particular, we see on Figure 11 that even if $P(c = manmade \mid v)$ and $P(c = natural \mid v)$ are estimated on the segmented image regions from the test set, there is an important ambiguity of the patches with respect to the two classes.

7.2 Gaussian Mixture Model soft assignment

Quantizing image regions into patches discard all the information about the distance of each particular local descriptor $s$to the corresponding patch cluster center v. It results in a compact representation that can be seen as a drastic simplification of the data. Two descriptors of highly similar local textures can

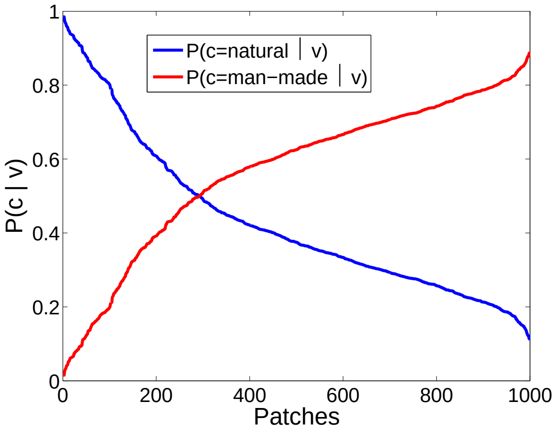


Figure 11: $P(c \mid v)$ for *man-made* and *natural* structures, estimated on the annotated patches from test images. The x axis is the patch indices ordered with decreasing $P(c = \mathit{natural} \mid v)$.

be assigned to different patches if their are close to the border between the two clusters. This intrinsic ambiguity of the quantization approach that can be questioned. In the previous example, knowing that the two regions were in fact similar could be beneficial.

One way to address this issue is to perform a soft clustering of the region features. Instead of attributing a single patch number to each local descriptor, we allow for multiple cluster assignments with membership probabilities, assuming that the region descriptors have been generated by a Gaussian Mixture Model (GMM) [2]. Given this soft clustering, we base the classification of image patches on the class likelihood ratio of their corresponding local descriptor s_j, given by:

$$LR(s_j) = \sum_{i=1}^{N_g} p(g_i \mid s_j) LR(g_i), \tag{18}$$

where N_g is the total number of Gaussian distributions in the GMM, $p(g_i \mid s_j)$ denotes the probability of the Gaussian g_i having generated the local descriptor s_j, and $LR(g_i)$ is the class likelihood ratio of the Gaussian g_i. Note that the empirical baseline based on the K-means hard clustering becomes a special case of Equation 18 when $p(g_i \mid s_j)$ equals 1 for one Gaussian component and 0 for others. The posterior probability $p(g_i \mid s_j)$ is computed as:

$$p(g_i \mid s_j) = \frac{p(s_j \mid g_i) p(g_i)}{p(s_j)}, \tag{19}$$

where $p(s_j \mid g_i)$, $p(g_i)$, and $p(s_j)$ relate to the standard GMM formulation. Each feature s_j is generated by a mixture of N_g Gaussian distributions, with the following likelihood given the estimated GMM mixture weights w, means μ, and standard deviations Σ:

$$p(s_j) = \sum_{i=1}^{N_g} p(s_j, g_i) = \sum_{i=1}^{N_g} p(g_i) p(s_j \mid g_i) = \sum_{i=1}^{N_g} w_i \mathcal{N}(s_j; \mu_i, \Sigma_i), \tag{20}$$

where $\mathcal{N}(s; \mu_i, \Sigma_i)$ is the Gaussian distribution of the component g_i. The class likelihood ratio of a Gaussian distribution is given by:

$$LR(g) = \frac{P(g|c = \text{man-made})}{P(g|c = \text{natural})}, \tag{21}$$

where $P(g|c)$ is estimated by the ratio of importance of that generating Gaussian distribution for each class in the labeled images.

8 Markov Random Field (MRF) regularization

The contextual modeling with latent aspects that we present in this paper can be conveniently integrated with traditional spatial regularization schemes. To investigate this we present the embedding of our contextual model within the MRF framework [14], though other schemes could be similarly employed [18, 19, 40].

Let us denote by S the set of sites s, and by $\mathcal{Q}$ the set of cliques of two elements associated with a second-order neighborhood system $\mathcal{G}$ defined over S. The patch classification can be classically formulated using the Maximum A Posteriori (MAP) criterion as the estimation of the label field $C = \{c_s, s \in S\}$ which is most likely to have produced the observation field $V = \{v_s, s \in S\}$. In our case, the set of sites is given by the set of interest points, the observations v_s take their value in the set of patches $\mathcal{V}$, and the labels c_s belong to the class set $\{man - made, natural\}$. Assuming that the observations are conditionally independent given the label field (i.e. $p(V|C) = \prod_s p(v_s|c_s)$), and that the label field is an MRF over the graph $(S, \mathcal{G})$, then due to the equivalence between MRF and Gibbs distribution ($p(x) = \frac{1}{Z}e^{-U(x)}$), the MAP formulation is equivalent to minimizing an energy function [14]

$$U(C,V) = \underbrace{\sum_{s\in S} V_1(c_s) + \sum_{\{t,r\}\in\mathcal{Q}} V_1'(c_t,c_r)}_{U_1(C)} + \underbrace{\sum_{s\in S} V_2(v_s,c_s)}_{U_2(C,V)}, \tag{22}$$

where U_1 is the regularization term which accounts for the prior spatial properties (homogeneity) of the label field, whose local potentials are defined by:

$$\begin{aligned} V_1(\text{man-made}) = \beta_p \;\text{ and }\; V_1(\text{natural}) = 0, \\ V_1'(c_t,c_r) = \beta \;\text{ if }\; c_t \neq c_r, \text{ and } V_1'(c_t,c_r) = 0 \;\text{ otherwise.} \end{aligned} \tag{23}$$

β is the cost of having neighbors with different labels, while β_p is a potential that will favor the man-made class label (if $\beta_p < 0$) or the natural one (if $\beta_p > 0$), and U_2 is the data-driven term for which the local potential are defined by:

$$V_2(v_s,c_s) = -\log(p(v_s|c_s)). \tag{24}$$

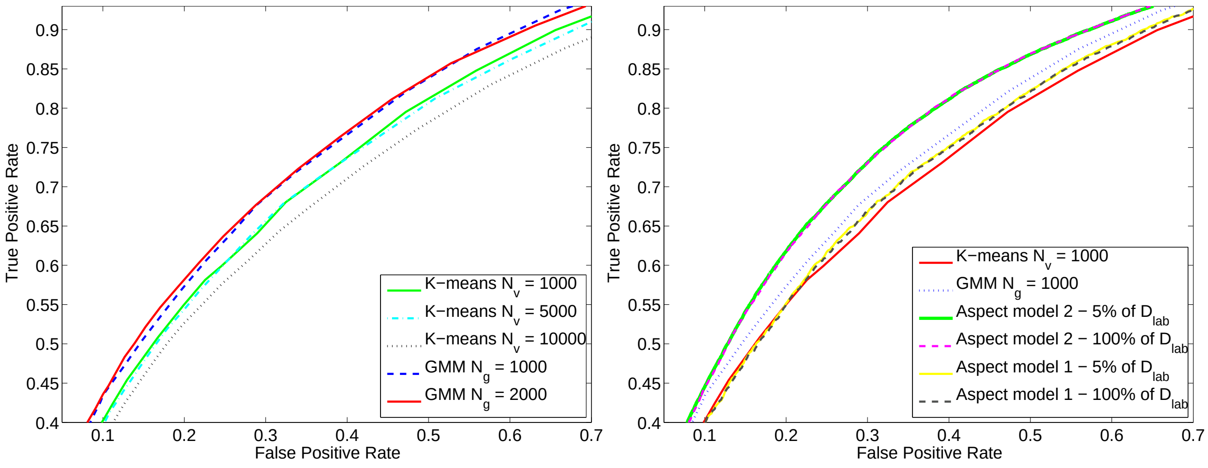
To implement the above regularization scheme, we need to specify a neighborhood system. Several alternatives could be employed, exploiting for instance the scale of the invariant detector (e.g. see [19]). Here we used a simpler scheme: two points t and r are defined to be neighbors if r is one of the $N_{\mathcal{N}}$ nearest neighbors of t, and vice-versa. For this set of experiments we defined the neighborhood to be constituted by the five nearest neighbors. Finally, in the experiments, the minimization of the energy function of Equation 22 was conducted using simulated annealing [21].

9 Experiments and discussion

We validate our proposed models on natural vs. man-made scene patch classification. In this Section, we present our experimental setup, show a detailed performance evaluation illustrated with the patch classification results on a few test images, and we finally study the result of integrating spatial regularization.

9.1 Experimental setup

Datasets: Three image subsets from the *Corel Stock Photo Library* were used in the experiments. The first set, $\mathcal{D}$, contains 6600 photos depicting mountains, forests, buildings, and cities. From this set, 6000 have no associated label, while the remaining subset $\mathcal{D}_{lab}$ is composed of 600 images, whose content mainly belonged to one of the two classes, which were hand-labeled with a single class label leading to approximately 300 images of each class. This labeling at the image level is used to quickly label the corresponding patches. $\mathcal{D}$ was used to construct the vocabulary and learn the aspect models, while $\mathcal{D}_{lab}$ was used, entirely or not, to estimate the patch likelihoods for each class. A third set $\mathcal{D}_{test}$, containing



(a) (b)

Figure 12: Comparison of the True Positive Rate vs. False Positive Rate curves for all patch classification methods, obtained by varying the likelihood ratio threshold T: (a) performance of the baseline methods for different numbers of K-means clusters and GMM components, (b) best baseline results compared to the aspect models with a vocabulary size of $N_v = 1000$ patches and $N_A = 20$ aspects.

485 images of man-made structures in natural landscapes hand-segmented with polygonal shapes to label the corresponding patches (Figure 1) was used to evaluate the methods.

Performance measure: The global performance of the algorithm was assessed using the True Positive Rate (TPR, number of positive regions correctly classified over the total number of positive descriptors), False Positive Rate (FPR, number of false positives over the total number of negative descriptors) and True Negative Rate (TNR=1-FPR), where man-made structure is the positive class. The FPR, TPR and TNR values vary with the threshold T applied for classification (see Equation 2).

Parameters: Results are reported with a vocabulary size ranging from 1000 to 10000 patches, a number of 1000 and 2000 GMM mixtures, and 20 aspects for aspect models 1 and 2.

9.2 Performance evaluation

Figure 12 displays the Receiver Operating Curve (TPR vs. FPR) of the empirical patch distribution baseline and the GMM baseline for various parameter settings (a), and gives a comparison between the baseline approaches with the best parameter settings with the two proposed aspect models (b). The ROC curves are obtained by varying the likelihood ratio threshold T, resulting in a different patch classification. The first observation relates to the influence of the patch vocabulary size, varied between 1000 and 10000 patches on Figure 12 (a), for the empirical patch distribution baseline. While no significant difference in performance is observed between the vocabulary of 1000 and 5000 patches, the performance decreases significantly for the 10000 patch vocabulary. This effect is somehow counter-intuitive since a higher granularity in the quantization allows to define a finer classification decision function. It can be explained by a higher level of noise in the estimation of the likelihood ratio, since the number of training images remains constant. In contrast, the GMM approach is more accurate, as it allows good likelihood ratio estimates while providing a finer feature space quantization through the soft assignment possibility. As in the two cases, no improvement is observed when using vocabulary sizes larger than 1000, we will use this number in the following (for the empirical patch distribution and the aspect models).

As can be seen on Figure 12 (b), the aspect model 1 performs slightly better than the empirical patch distribution baseline, for all vocabulary sizes. However, the GMM baseline improves both the empirical patch distribution baseline and the aspect model 1 classification performance. The GMM approach is therefore the best image independent patch classification approach. Aspect model 2 outperforms significantly all other methods, proving the advantage of an image dependent patch classification. Interestingly, the aspect models does not need 100% of the 600 labeled images for a good classification performance. We can observe on Figure 12 that the same patch classification performance is achieved when using only 5% of the labeled images (30 images) required to estimate the class-conditional aspect likelihood $P(z|c)$.

To further validate our approach, Table 1 reports the Half-Total-Recognition Rate (HTRR) measured by 10-fold cross-validation. For each of the folds, 90% of the test data $\mathcal{D}_{test}$ is used to estimate the likelihood threshold T_{EER} leading to Equal Error Rate (EER, obtained when TPR=TNR) on this data. This threshold is then applied on the remaining 10% (unseen images) of $\mathcal{D}_{test}$, from which the HTRR (HTRR=(TPR+TNR)/2) is computed. This table shows that the ranking observed on the ROC curve is clearly maintained, and that aspect model 2 results in a 7.5% performance relative increase w.r.t. the baseline approach.

As mentioned in Section 6, aspect model 1 and the empirical distribution method (GMM and K-means based) assign specific patches to the man-made or natural class independently of the actual image in which those patches occur. This sets a common limit on the maximum performance of both systems, which is referred here as the *ideal case.* This limit is given by attributing to each patch the class label corresponding to the class in which that patch occurs the most in the *test data.* On our data, this *ideal case* corresponds to a HTRR of 71.0% for the 1000 patches vocabulary, showing the advantage of an image dependent patch classification method.

In order to have a chance of performing better than the *ideal case*, patches must be labeled differently depending on the specific image that is being segmented. Aspect model 2 switches patch class labels according to the contextual information gathered through the identification of image-specific latent aspects. In our data, successful class label switching occurs at least once for 727 out of the 1000 patches in our vocabulary.

9.3 Patch classification examples

The impact of the contextual model can also be observed on individual images. Figure 13 displays classification examples of man-made image patches, where likelihood thresholds were estimated at EER value. As can be seen, aspect model 2 improves the classification results with respect to the two other methods in two different ways. On one hand, in the first three examples, aspect model 2 increases the precision of the man-made patch classification, producing a slight decrease in the corresponding recall. On the other hand, the fourth example shows aspect model 2 producing a higher recall of man-made patches while maintaining a stable precision. In the fifth example, the occurrence of a strong context causes the whole image to be taken as natural a scene, also improving the total patch classification.

In Figure 14, five more examples of patch classification are shown. The first three rows illustrate *natural* image context examples that are correctly grasped by aspect model 2. The fourth row shows a correctly estimated *man-made* context that leads to an improved classification of patches for aspect model 2. In the fifth example, however, the overestimation of the *man-made* related aspects leads to patches that are dominantly classified as *man-made.* Nevertheless, overall, as indicated in Figure 12 and Table 1, the introduction of context by co-occurrence is beneficial.

9.4 Effects of the Markov Random Field regularization

We investigate the impact of the combination with spatial regularization on the task of patch classification. The level of regularization is defined by β (a larger value implies a larger effect). The regularization is conducted by starting at the Equal Error Rate point, as defined in the 10-fold cross-validation experiments described in preceding Section. More precisely, for each of the folds, the threshold T_{EER} is used to set the prior on the labels by setting $\beta_p = -\log(T_{EER})$. Thus, in the experiments, when $\beta = 0$ (i.e. no spatial regularization is enforced), we obtain the same results as in Table 1. In Figure 15 we see that the best patch classification performance corresponds to an HTRR of 73.1% and a β of 0.35 with the empirical modeling, and an HTTR of 76.3% for a β of 0.2 and aspect model 2. This latter value of β is chosen for all the MRF illustrations reported in Figure 16 and 17.

The inclusion of the MRF relaxation boosted the performance of both aspect model 2 and empirical

	Emp. dist. K-means	Emp. dist. GMM	Aspect model 1	Aspect model 2
HTRR	67.5	69.7	68.5	72.4

Table 1: Half Total Recognition Rate (in percent).

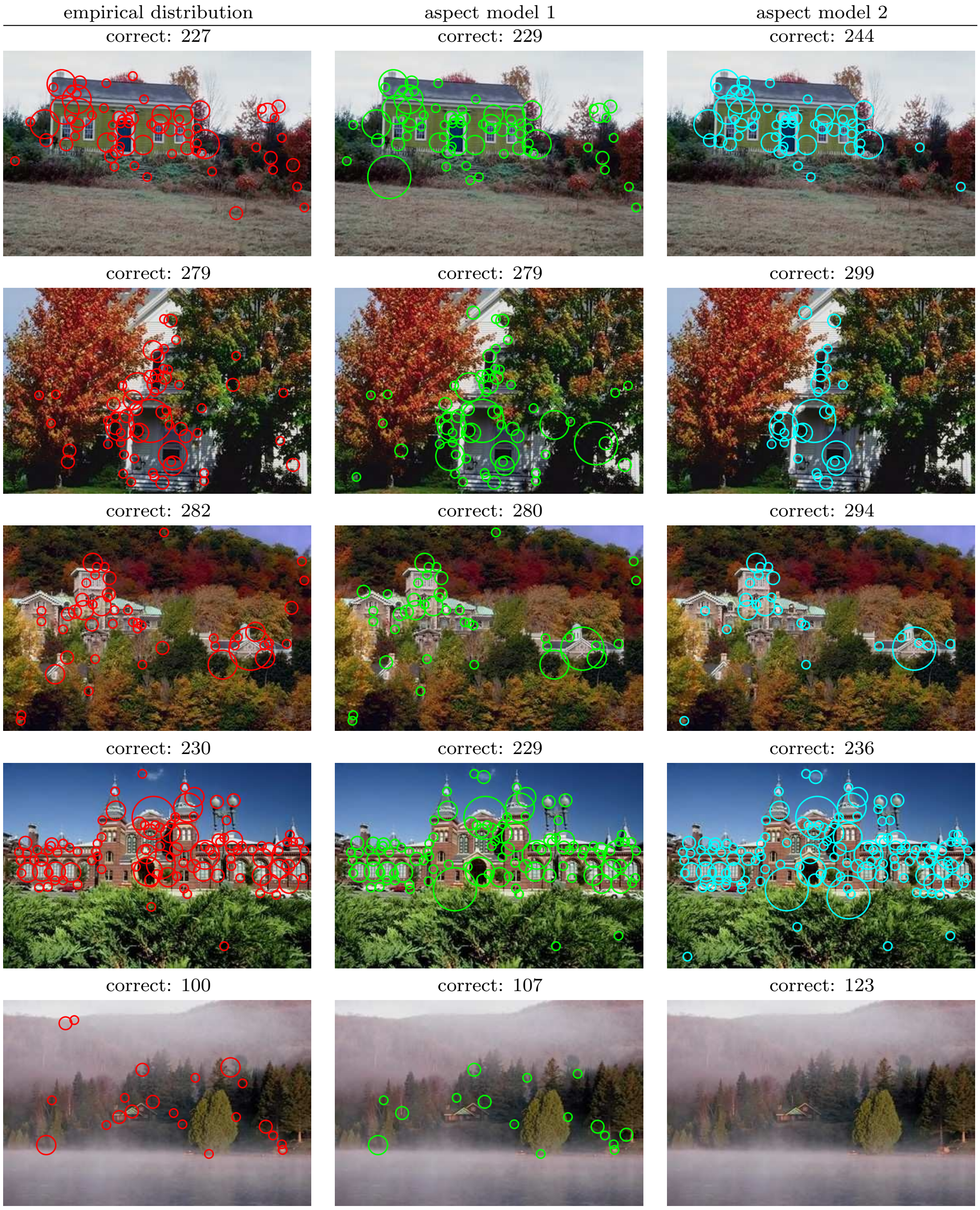


Figure 13: Image patch classification examples at T_{EER}. Results provided by: first column, K-means empirical distribution; second column, aspect model 1; third column, aspect model 2. The total number of correctly classified patches (man-made + natural) is given per image. The five rows illustrate cases where aspect model 2 outperforms the other approaches. In the fifth row, an extreme example of a strong natural context that is correctly identified by aspect model 2 leads to the classification of all regions as natural (though some should be labeled as man-made).

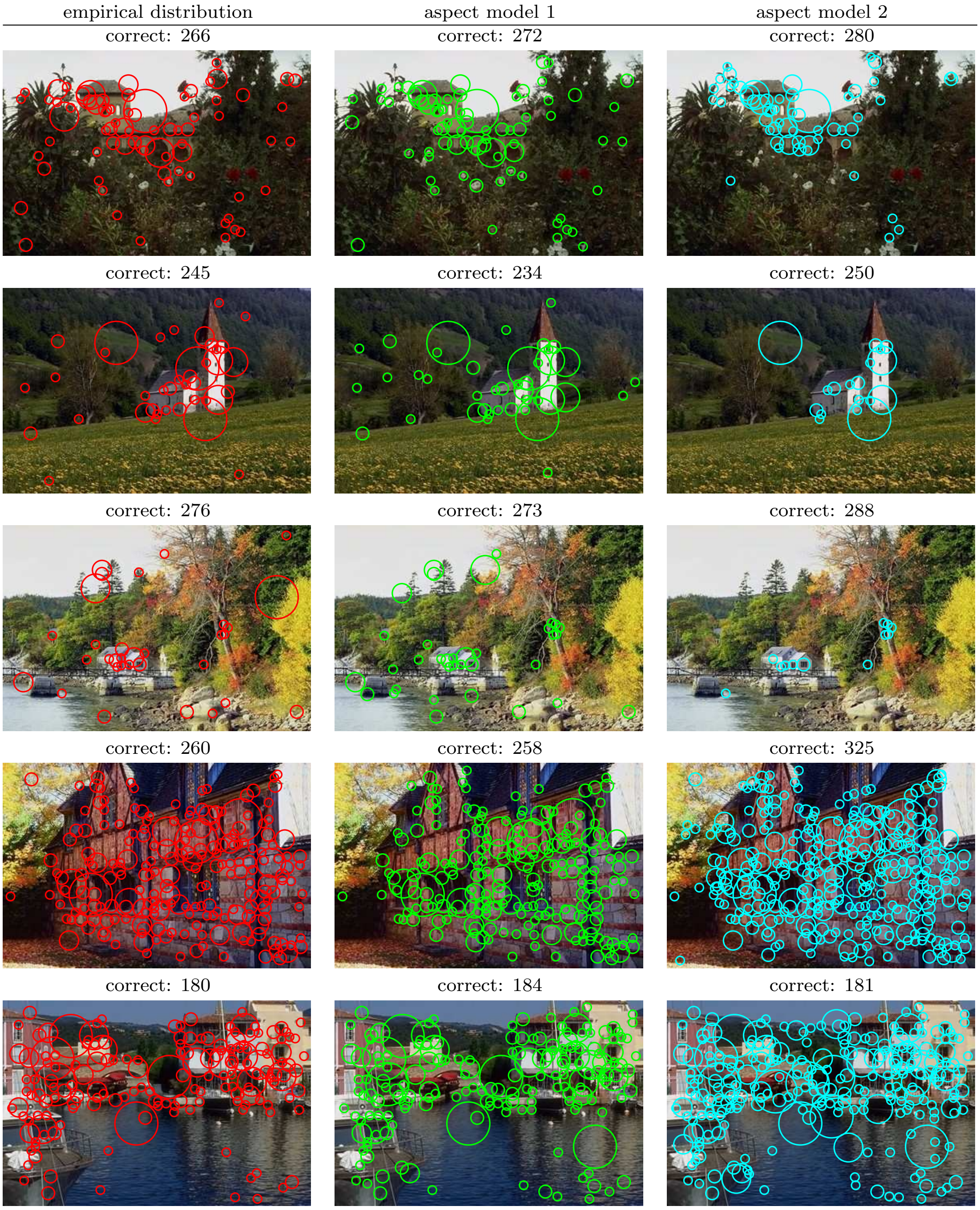


Figure 14: Image patch classification examples at T_{EER}. Results provided by: first column, K-means empirical distribution; second column, aspect model 1; third column, aspect model 2. The first three rows illustrate the case of a correctly identified marked *natural* image context by aspect model 2, resulting in a more accurate patch classification as compared to aspect model 1 and empirical distribution. The fourth row shows a correctly identified marked *man-made* image context by aspect model 2,with an improved number of correctly classified points. The last row shows the confusion in patch classification, when the context is not correctly identified (in this case, overestimated) by aspect model 2.

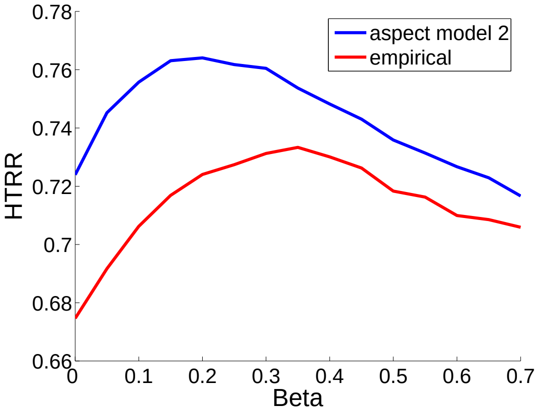


Figure 15: The evolution of the Half Total Recognition Rate for different β values for the MRF regularization.

distribution. However, it is important to point out that aspect model 2 still outperforms the empirical distribution model, though the boosting beneficiated most to the empirical distribution modeling. This was to be expected, as aspect model 2 was already capturing some of the contextual information that the spatial regularization can provide (notice also that the maximum is achieved for a smaller value of β in aspect model 2).

Besides obtaining an increase of the HTRR value, we can visually notice a better spatial coherence of the patch classification, as can be seen in Figure 16 and 17. We can observe in the images that the MRF relaxation process reduces the occurrence of isolated points, and tends to increase the density of points within segmented regions. We show on the last row of Figure 16 that as can be expected when using prior modeling, on certain occasions the MRF step can over-regularize the patch classification, causing the attribution of only one label to the whole image.

10 Conclusion and future work

In this paper, we proposed computational models to perform contextual regional classification of images. These models enable us to exploit a different form of visual context, based on the co-occurrence analysis of patches in the whole image rather than on the more traditional spatial relationships. Patch co-occurrence is summarized into aspects models, whose relevance is estimated for any new image, and used to evaluate class-dependent patch likelihoods. These models have been tested and validated on a man-made vs. natural scene image patch classification task. One model has clearly shown to help in disambiguating polysemic patches based on the context they appear in. Producing satisfactory classification results, it outperforms state-of-the-art likelihood ratio methods [9], even when using soft assignment techniques.

Moreover, we investigated the use of Markov Random Field models to introduce spatial coherence in the final classification and show that the two types of context models can be integrated successfully. This additional information enables to overcome some patch classification errors from the likelihood ratio and aspect models methods, increasing the final performance.

While the results presented here are encouraging, this task is complex, and there is a need for further improvements. Logical extensions would be the introduction of other sources of contextual information like color or scale and other forms of integration of spatial contextual information.

Acknowledgments

This work was funded by the Swiss National Science Foundation through the MULTI project and the Swiss National Center of Competence in Research on Interactive Multimodal Information Management (NCCR (IM)2).

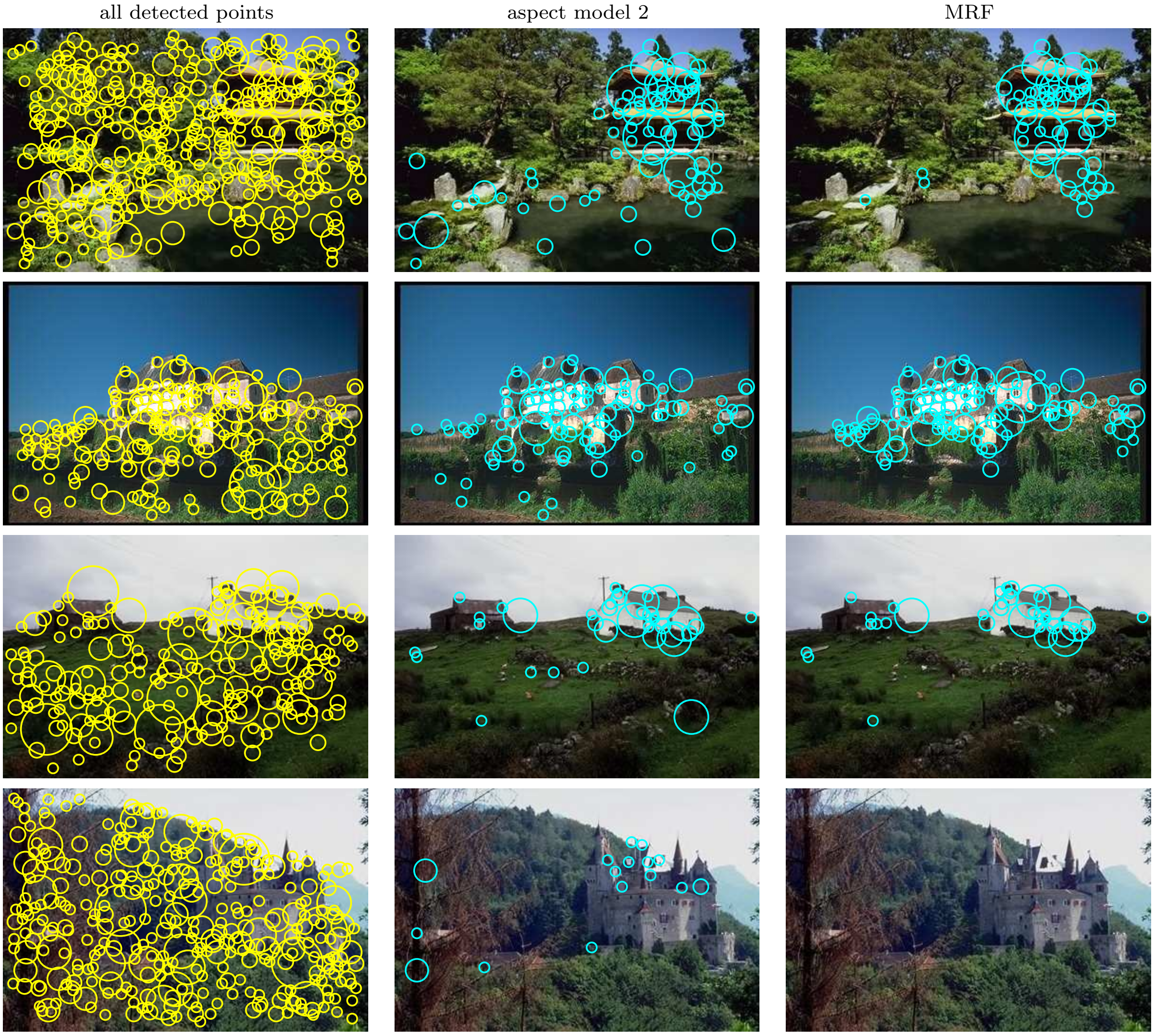


Figure 16: Effect of the MRF regularization on the man-made patch classification. The first three rows illustrate the benefit of the MRF regularization where wrongly classified isolated patches are removed. The last row shows the deletion of all man-made classified patches from an image when natural patches dominate the scene.

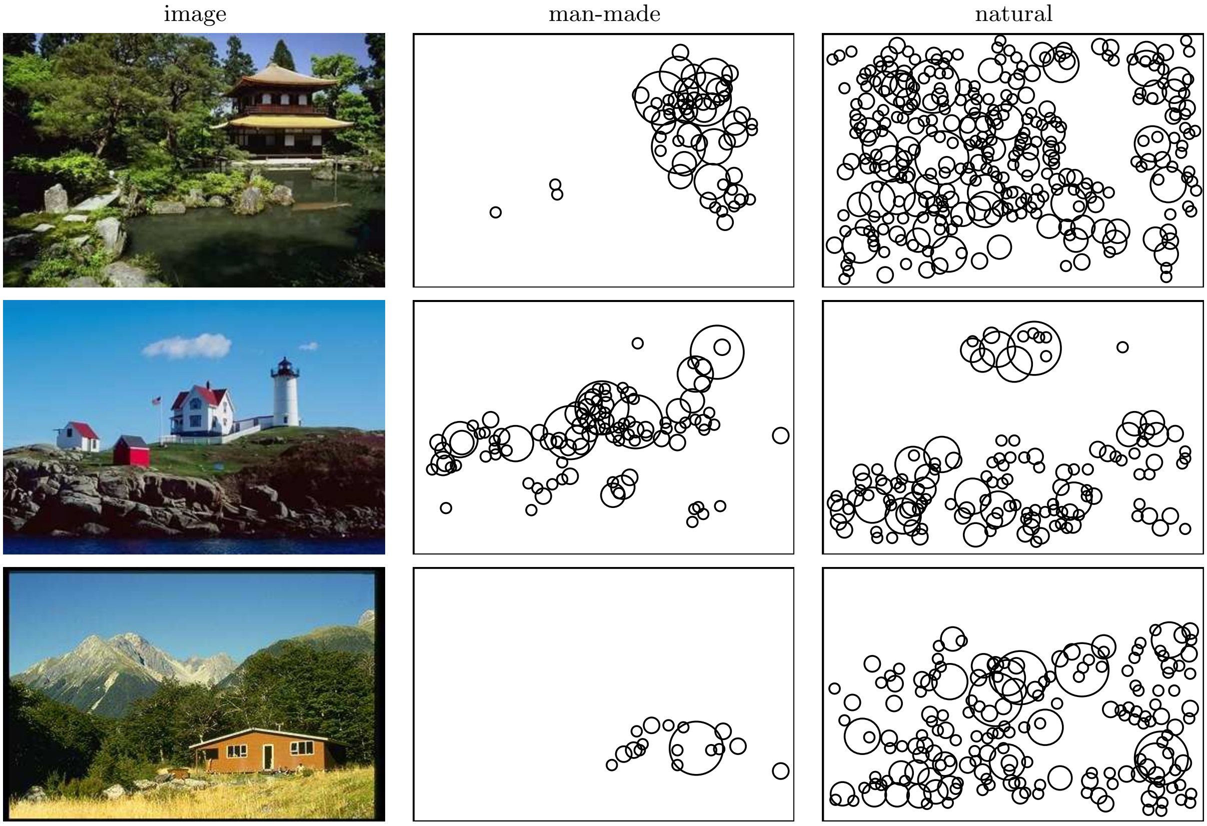


Figure 17: Three other examples that illustrate the final patch classification obtained with aspect model 2 and MRF regularization. The display is different than in previous figures to avoid image clutter.

References

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

[2] C. Bishop. *Neural Networks for Pattern Recognition.* Oxford University, 1995.

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

[4] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification via plsa. In *ECCV (4)*, pages 517–530, 2006.

[5] W. Buntine. Variational extensions to EM and multinomial PCA. In *Proc. of Europ. Conf. on Machine Learning*, Helsinki, Aug. 2002.

[6] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proc. of the IEEE International Conference in Computer Vision*, Rio de Janeiro, Oct. 2007.

[7] Chad Carson, Megan Thomas, Serge Belongie, Joseph M. Hellerstein, and Jitendra Malik. Blobworld: A system for region-based image indexing and retrieval. In *Third International Conference on Visual Information Systems.* Springer, 1999.

[8] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proc. of the European Conference on Computer Vision*, Prague, May 2004.

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

[10] P. Duygulu, K. Barnard, N. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of the IEEE Europ. Conf. on Computer Vision*, Copenhagen, May 2002.

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

[12] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *Proc. of ICCV 2005*, Beijing, Oct. 2005.

[13] Y. Zhang G. Wang and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. In *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition*, New York, Jun. 2006.

[14] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

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

[16] M. Kumar, P. Torr, and A. Zisserman. Obj cut. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, Jun. 2005.

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

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

[19] S. Lazebnik, C. Schmid, and J. Ponce. Affine-invariant local descriptors and neighborhood statistics for texture recognition. In *Proc. of Int. Conf. on Computer Vision*, Nice, Oct. 2003.

[20] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Proc. of the ECCV workshop on statistical learning in computer vision*, Prague, May 2004.

[21] S. Z. Li. *Markov Random Field Modeling in Computer Vision.* Springer, 1995.

[22] D. Liu and T. Chen. Background cutout with automatic object discovery. In *Proc. of the IEEE International Conference on Image Processing*, San Antonio, Sep. 2007.

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

[24] M. Marszaek and C. Schmid. Spatial weighting for bag-of-features. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Washington, 2006.

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

[26] 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*, 2005.

[27] F. Monay, P. Quelhas, D. Gatica-Perez, and J.-M. Odobez. Constructing visual models with a latent space approach. In *Subspace, Latent Structure and Feature Selection*, volume 3940/2006 of *Lecture Notes in Computer Science*, pages 115–126, 2006.

[28] F. Monay, P. Quelhas, D. Gatica-Perez, and J.-M. Odobez. Integrating co-occurrence and spatial contexts on patch-based scene segmentation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, New York, Jun. 2006.

[29] K. Murphy, A. Torralba, and W. Freeman. Using the forest to see the trees: A graphical model relating features, objects and scenes. In *Proc. of Neural Information Processing Systems*, Vancouver, Dec. 2003.

[30] N. R. Pal and S. K. Pal. A review on image segmentation techniques. *Pattern recognition*, 26:1277–1294, 1993.

[31] P. Quelhas, F. Monay, J.-M Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29:1575–1589, 2007.

[32] 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 the IEEE Int. Conf. on Computer Vision*, Beijing, Oct. 2005.

[33] Pedro Quelhas and Jean-Marc Odobez. Multi-level local descriptor quantization for bag-of-visterms image representation. In *Proceedings of CIVR*, pages 242–249, 2007.

[34] B. Russell, W. Freeman, A. Efros Alexei, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, New York, Jun. 2006.

[35] J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition*, San Juan, Jun. 1997.

[36] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 1–15, 2006.

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

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

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

[40] J. J. Verbeek and B. Triggs. Region classification with markov field aspect models. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[41] 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.

[42] Julia Vogel and Bernt Schiele. A semantic typicality measure for natural scene categorization. In *Proc. of the Pattern Recognition Symposium DAGM'04*, Tubingen, September 2004.

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