

Latent Mixture Vocabularies for Object Categorization and Segmentation

Diane Larlus, Frédéric Jurie

INRIA Rhône-Alpes
Fax: +33 476 61 54 54
655 avenue de l'Europe
Montbonnot
F-38 334 Saint Ismier Cedex, France

Abstract

The *visual vocabulary* is an intermediate level representation which has been proved to be very powerful for addressing object categorization problems. It is generally built by vector quantizing a set of local image descriptors, independently of the object model used for categorizing images. We propose here to embed the visual vocabulary creation within the object model construction, allowing to make it more suited for object class discrimination and therefore for object categorization. We also show that the model can be adapted to perform object level segmentation task, without needing any shape model, making the approach very adapted to high intra-class varying objects.

Key words: Object categorization, object segmentation, visual vocabulary creation

1 Introduction

Object categorization is an important task in computer vision, which has received a lot of attention over the last three years [4,6,9,13,17,22–24,26]. This problem is challenging because pose and illumination changes, scale variations as well as occlusions and intra-class variability can make two images of the same class very different.

Methods which were first proposed, like QBIC [8], were based on feature vectors encoding global properties (color, shape or texture, etc.) of images.

Email address: `firstname.name@inrialpes.fr` (Diane Larlus, Frédéric Jurie).
URL: `http://lear.inrialpes.fr` (Diane Larlus, Frédéric Jurie).

However, except for simple visual classes, it is difficult to reliably link image semantics with global representations. The rationale behind this difficulty is that objects of interest, or the visual information providing evidence for categories, often constitute small fractions of images. They can therefore barely be detected in global signatures.

More recently, methods based on the analysis of local information such as image patches [4,24] have been shown to outperform global methods. The challenge becomes building class models capable of extracting semantics from loose sets of image patches, even when only a few of them are informative.

Finding class models that are invariant enough to cope with intra-category variations and discriminative enough to distinguish between classes is the key issue of object categorization.

1.1 Related work

Very efficient statistical models have been used to address this problem; they were often inspired by text analysis. After building a *visual vocabulary*, images can be processed as sets of visual words and frameworks used for categorizing text become applicable. One of the most successful models is the *bag-of-features* model, first applied to image categorization by [4] and [24], and later extended by many other authors [17,26]. Images are simply modeled by measuring frequencies of unordered sets of visual words, encoded as histograms.

The bag-of-features strategy inspired more complex models, like probabilistic Latent Semantic Analysis (pLSA) [12], or its Bayesian form Latent Dirichlet Allocation (LDA) [2]. These models have recently been applied to object categorization [6,7,22,23,25]. They consider visual words as generated from latent aspects (or *topics*). The model expresses images as combinations of specific distributions of topics.

All of these methods require images to be translated into visual words, this intermediate representation linking concepts with image pixels, by a distinct process. Visual vocabularies generally result from a quantization process: a collection of visual features (such as patches) are sampled on a set of training images, encoded into a convenient representation (like the popular SIFT representation [19]), and vector quantized by a clustering algorithm.

Several combinations of patch detectors, visual descriptors and clustering algorithms have been proposed in the past. The most popular way consists in detecting interest points and clustering their SIFT representation with k-means, as originally proposed by [4,24]. Agglomerative techniques [18] or mean-shift

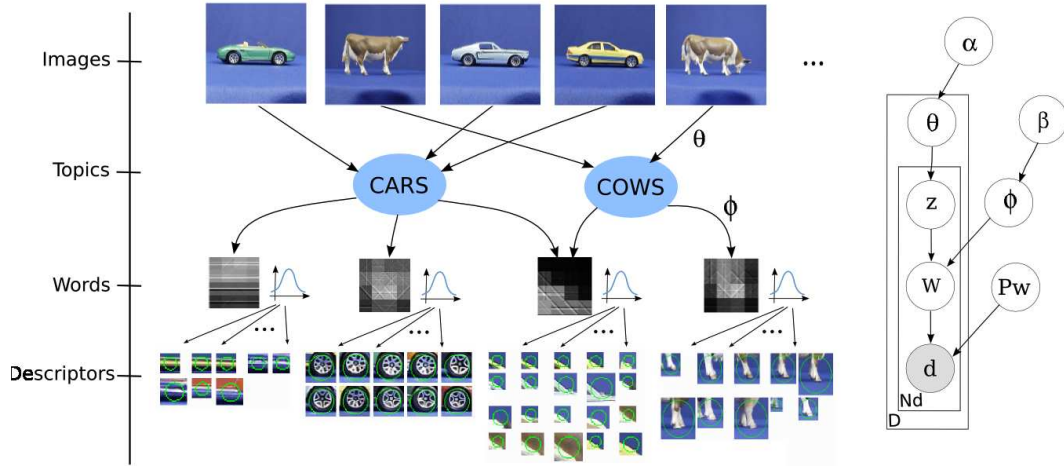


Fig. 1. Overview of the latent mixture vocabulary model, and the corresponding graphical model representation.

based approaches [13] have also been used for their capability of dealing with unbalanced clusters. In both cases, histograms can be built by assigning each feature vector to its closest centroid.

Whatever algorithm is used, for all of the previously mentioned approaches, building the visual vocabulary is a distinct preprocessing stage and not a component of the model. However, contrary to text, visual vocabulary is an artificial concept, not uniquely defined, but on which image representation and then classification performances strongly depend. The efficiency of vocabularies estimated without any regard for the classification task nor with the image modeling process should be questioned.

In [26], authors cope with this issue and suggest to build a compact and more discriminative vocabulary by pair-wise merging of visual words, from an initial large vocabulary. However, if two distinct visual words are initially grouped in the same cluster, they can not be separated later.

This idea of building adapted vocabularies has also been explored recently by Perronnin *et al.* [20]. They address this issue by combining a universal vocabulary with class specific vocabularies. The universal vocabulary describes the visual content of all the considered classes while a class specific vocabulary is obtained by adapting the universal vocabulary to a class using specific data. This combination of universal and specific approaches constitutes an interesting contribution to the computation of adapted vocabularies. However, these specialized vocabularies are designed to emphasize differences between a mean histogram and a class specific histogram, but not to emphasize differences between classes. If two classes are visually close, there is no guarantee that some words will help to distinguish one from another.

1.2 Overview of the proposed approach

The approach proposed in this article tries to go one step further in the aim of producing visual vocabularies adapted to the classification task. Inspired by [22,23], we propose a generative model based on latent aspects for explaining images at feature descriptors level. Instead of using a vocabulary computed in a preprocessing stage, the visual vocabulary is a built-in component of the model, learned simultaneously with other parameters. Indeed, we consider images as distributions over *topics*, topics as distributions over *words* and words as Gaussian mixture densities over visual descriptors (see Figure 1 for an illustration of the model).

Dirichlet priors on topic and word distributions tense to produce a few class specific visual words and more generic words shared between classes. Interestingly, our model can be learned without any supervision, whereas we argue later that a little supervision can make the estimation more stable.

The organization of the paper is as follows: section 2 presents our latent mixture vocabularies model, the way to estimate its parameters and the way to use it in a classification framework. Section 3 presents an extension of this model allowing to segment images. Section 5 contains experimental results for both the classification and the segmentation tasks. At last, section 6 concludes and give a few perspectives.

2 Modeling local appearance statistics

2.1 Model description

Images are considered as unordered sets of visual descriptors, found using an interest point detector or uniformly sampled on images¹. In this article, visual descriptors are SIFT vectors in a 128-dimensional space, but other descriptors could be used. Position and scale of the descriptors in the image are not used.

We use a simplified form of the Gaussian-Multinomial LDA model (GM-LDA) [1], which is a latent variable model that allows visual descriptors to be allocated repeatedly in images. Visual descriptors come from two underlying factors, denoted *topics* and *visual words*. Images are modeled as combinations of T possible topics which themselves produce N visual words, while words are Gaussian distributions over the SIFT descriptor space. Topic distributions

¹ In practice, according to [13] and [26] we pick patches on a regular grid at multiple scales but other strategies could be used.

over words (ϕ) are sampled from a Dirichlet distribution of parameter β . This property is shared between all images.

Modeling image I with our model assumes it is built according to the following generative process:

- (1) sample $\theta \sim Dir(\alpha)$, where $Dir(\alpha)$ is a Dirichlet distribution with hyper-parameter α , providing a distribution over the latent *topic* factors, specific for this image.
- (2) For each image descriptor d_i ,
 - (a) sample a topic z_i from the multinomial distribution with parameter θ , $z_i \sim Mult(\theta)$.
 - (b) sample a visual word w_i conditional on z_i from the multinomial distribution with parameter ϕ_{z_i} , $w_i \sim Mult(\phi_{z_i})$. This sampling does not depend on the image anymore, only on the previously sampled topic.
 - (c) finally, sample a visual descriptor d_i conditional on w_i , $d_i \sim \mathcal{N}(P_{w_i})$, where $\mathcal{N}(P_{w_i})$ denotes the Gaussian distribution with parameter P_{w_i} . This sampling still does not depend neither on the image nor on the topic, but only on the visual word.

The resulting distribution on the set d of visual descriptors belonging to image I is given as follows:

$$p(d|P, \phi, \alpha, \beta, I) = \int \prod_{d_i \in I} \sum_{j=1}^N \sum_{k=1}^T p(d_i|w_j, P) p(w_j|z_k, \phi) p(z_k|\theta) p(\theta|\alpha) d\theta \quad (1)$$

Compared to [6,7,22,23] our model has an extra layer responsible for the generation of visual descriptors conditional to visual words. This layer is the key part of our model as it allows to learn the visual vocabulary.

The graphical model representation can be found in Figure 1.

2.2 Model estimation

Hyper-parameters α and β play an important role as they allow to control how sparse and therefore specialized topics and visual words distributions can be. This is why, according to [10] we prefer not to estimate them and use fixed Dirichlet priors. Learning the model consists of a likelihood maximization and is done by estimating the optimal parameters ϕ and P , for a given set of images.

Since the integral in equation (1) makes the direct optimization of the likeli-

hood intractable, we estimate variables of interest by an approximate iterative technique called Gibbs Sampling. It is a special case of Markov chain Monte Carlo where a Markov chain is constructed to converge to the target distribution. The next state of that chain is reached by sequentially sampling all variables from their distribution when conditioned on the current values of all other variables and the data.

In our model we can use the fact that priors (α and β) are conjugate to the multinomial distributions ϕ and θ , and then consider the posterior distribution over the assignments of words to topics $p(z|w)$. A complete justification can be found in [10] for the LDA model.

The estimation process is done by sampling the distributions $p(z_i|z_{-i}, w)$ and $p(w_i|d_i, w_{-i}, z, P)$ for all observations d_i where z_{-i} represents all z values except z_i . The first distribution is obtained using counts over previously sampled variables,

$$p(z_i = j|z_{-i}, w) \propto \frac{(n_{-i,j}^{w_i} + \beta)(n_{-i,j}^I + \alpha)}{(n_{-i,j}^{(\cdot)} + W\beta)(n_{-i,\cdot}^I + T\alpha)} \quad (2)$$

with $n_{-i,j}^w$ being the number of times word w has been assigned to topic j excluding the currently considered observation i , and $n_{-i,j}^I$ being the number of times a word from image I has been assigned to topic j . $n_{-i,j}^{(\cdot)}$ represents the total number of words assigned to topic j and $n_{-i,\cdot}^I$ is the number of observations in image I excluding descriptor d_i .

The second distribution comes from $p(w_i|d_i, w_{-i}, z, P) \propto p(d_i|w_i, P)p(w_i|t_i)$ where $p(w_i|t_i)$ corresponds to ϕ_j^w for $w_i = w$ and $t_i = j$ which can be obtained by $\frac{n_{-i,j}^w + \beta}{n_{-i,j}^{(\cdot)} + W\beta}$. The distribution $p(d_i|w_i, P)$ is equal to $\mathcal{N}(P_{w_i})$ following the model. P corresponds to the Gaussian mixture parameters which describe the words and is re-estimated at each iteration using standard sampling techniques (Gaussian distribution for the means and Wishart distribution for the covariances).

This iterative process is initialized using equiprobable distribution over topics, and k-means is used to create initial visual words.

The parameters of the model can be estimated without any need for supervision, i.e. using unlabeled training images. It is expected that for a given image, if we marginalize the probability given the descriptors belonging to that image, $p(\theta, \phi, P|d)$ over ϕ and P , $p(\theta|d)$ will have modes correlated with true classes, allowing to have class specific visual words. Unfortunately, we experimentally observed that it was not the case when images were cluttered and when objects occupy only a small fraction of the image. In this case, the

model is better described using topics which are not correlated to the classes we are interested in. We observed that this behavior can be avoided by adding supervision, assuming topics are known (derived from class labels) for a few training images. In all cases, we experimentally observed that this kind of supervision leads to a more accurate description of the classes, and improves final classification performances a lot.

2.3 *Classifying images*

The previous sections described the model we propose and explain how it can be learned from training data. We are now going to see how to use it for classification tasks, our ultimate goal. We investigate and compare different possible tracks, making different uses of the learned parameters.

2.3.1 *Image classification based on topic likelihood*

Topics are designed for being good representative of document contents. It is therefore natural to try to use them in the classification rule; we embed them in Maximum Likelihood (ML) criterion². The most straightforward way to implement this rule is to set the number of topics equal to the number of object classes and to assume that the class probability is equivalent to the topic probability, given an image. For example, if class C_i is represented by topic z_i in image I , we have $p(C_i|I) = p(z_i|I) = \theta_i$. We experimentally observed that the Markov chain generated by the Gibbs sampler for θ tends to converge quickly towards sharp and stable modes. In practice we use the expectation of the posterior distribution θ_i and this expectation is approximated by the last samples for θ_i . We denote this rule the TOPIC-BAYES classifier.

2.3.2 *Topic based SVM classifier*

However with more topics than classes the TOPIC-BAYES rule cannot be applied anymore. We adopted the more general classification scheme proposed in [22]. This scheme consists in training a classifier on the latent variables associated with each image. This cannot be directly done with our Gaussian-LDA model which does not explicitly estimates numerical values for latent variables but probability densities. However, as stated before, the sampler obtains stable modes and we assign the values corresponding to these modes to each image. An SVM classifier is trained on these values. We call this classifier the TOPIC-SVM classifier.

² We do not consider here the Maximum A Posteriori criterion as prior on classes are generally not available.

2.3.3 Bag-of-features based SVM classifier

Finally, instead of classifying images from their topic distributions, images can also be classified according to their visual words statistics, as it was done in the original bag-of-features approach. We use the same approach but instead of using a simple quantization of the descriptors to build the visual vocabulary, the words learned by our model are used. Comparing bag-of-features with classification from topics for the same model is an interesting issue. We denote this classification rule as the LDA-VOC-SVM rule.

3 Object level segmentation

The model proposed in the previous section describes images as a loose collection of visual concepts, the topics. However, once the topics have been identified, it becomes possible to establish connections between these topics and visual patches through visual words. It therefore opens the door to segmentation applications, if this connexion can be extended to image pixels.

In this section we show how the previous model can be adapted to image segmentation and labeling problems. These problems consist in separating or grouping image pixels into consistent parts, expected to be elements that humans consider as individual objects or distinct object parts.

This problem received a huge amount of attention in the past, and was originally addressed as an unsupervised problem. Many different methods have been developed, using various image properties such as color, texture, edges, motion, etc. [11]. It eventually turned out that image segmentation and image understanding were two closely related problems which cannot be solved independently.

After being abandoned for a while, image segmentation came back into favor recently, taking advantage of recent advances in machine learning.

The goal addressed here is the segmentation of objects belonging to a given category (the so-called *figure-ground segmentation* problem) assuming the category is defined by a set of training images. This is illustrated Figure 2 for a very challenging category, the “bicycle” category. The overall objective is to classify image pixels as being ‘*figure*’ or ‘*ground*’. Objects can have any size and any position in the image. They can occur with widely varying orientations and appearances.

In such conditions, object segmentation is intimately linked to object detection and recognition. Indeed, segmenting objects requires to learn object models

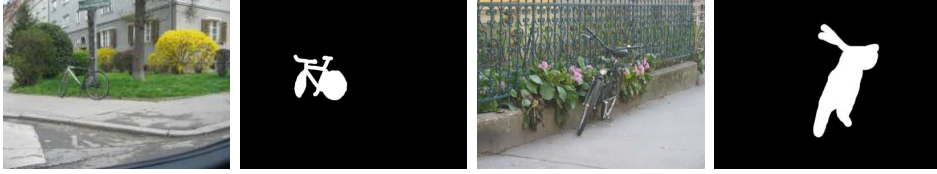


Fig. 2. The problem of figure/ground segmentation can be summarized like this: here we have two images of the “bike” category, with their corresponding hand-made segmentation masks. Our goal is to design algorithms capable of computing automatically this segmentation, after a training stage where the class model is learned.

from training images, as well as to search for occurrences of these models in images.

In this article we focus on difficult real-condition images where the objects can present extreme appearance variations (see Figure 2).

The proposed segmentation method is inspired by several related recent works, summarized below.

Leibe and Schiele [18] were among the first authors proposing to learn how to segment objects. Their contribution is before all a method for categorizing unseen objects in difficult real-world scenes. The method generates object hypotheses, without prior segmentation, that can be exploited to obtain a category-specific figure-ground segmentation. Training images are used to build a visual vocabulary of interest points, containing information about their relative positions as well as their corresponding segmentation masks.

Borenstein et al. [3] used the same idea of selecting informative patches from training images and using their corresponding segmentation masks in order to find object regions on new unseen images. They combined bottom-up and top-down approaches into a single process. The top-down approach uses an object representation learned from examples to detect an object in a new image and provides an approximation to its figure-ground segmentation. The bottom-up approach uses image-based criteria to define coherent groups of pixels that are likely to belong together to either the figure or the background. The combination of both approaches provides a final segmentation that benefits from both.

Several approaches proposed to use Conditional Random Field (CRF) for part-based detection [21] or segmentation [14]. Kumar et al. [14] proposed another methodology for combining top-down and bottom-up cues with CRFs. They combined CRFs and pictorial structures (PS). In the standard CRFs based segmentation, a contrast term favors grouping in the same label pixels with similar colors. However, due to the lack of a shape model, these methods do not work well for automatic segmentation. The PS provides good priors to CRFs

for specific shapes. In the standard CRFs based segmentation, a contrast term favors grouping in the same label pixels with similar colors. However, due to the lack of a shape model, these methods do not work well for automatic segmentation.

Kumar and Hebert [15] introduced the notion of Discriminative Random Fields (DRFs) by exploiting probabilistic discriminative models instead of the generative models generally used in the MRF framework.

None of the previously mentioned approaches is able to cope with occlusion. Win and Shotton [27] were the first to specifically addressing this problem using an enhanced CRF. The main contribution of their approach is the introduction of asymmetric pairwise potentials. This allows the relative layout (above / below / left / right) of parts to be modeled, as well as the propagation of long-range spatial constraints using only local pairwise interactions.

3.1 Overview of the method

Our approach shares many common features with the previously mentioned approaches. Firstly, as for all of these previous works, it combines bottom-up and top-down strategies.

The bottom-up process consists in sampling visual features (patches) and quantifying their representation into the previously mentioned set of visual words. From this stage, images are seen as sets of visual words occurrences. As the labeling process (i.e. the segmentation process) assigns figure/ground labels to patches, the pixel level segmentation requires an additional process, responsible for combining labels carried by patches into pixel hypotheses. The top-down process embeds object models and uses them to obtain a global coherence, by combining local information provided by the bottom-up process.

Most of the models previously used in this context cannot be used here, because of the strong variation of object’s appearance. Geometric models such as the Pictorial Structure [14] or the Implicit Shape Model [18] would require a huge number of training images in case of complex object categories like “bicycles” in order to capture the large variability of appearance. Approaches based on characteristic edge patches [3] are only usable when object outlines are stable enough. As a consequence, it appears that a more flexible and adapted model is required to address such categories.

The latent aspect based framework we described in the previous section seems to be appealing in the context of images segmentation for several reasons. First because object appearances (topics) can be automatically discovered and learned, limiting the amount of supervision required. Second, the flexibil-

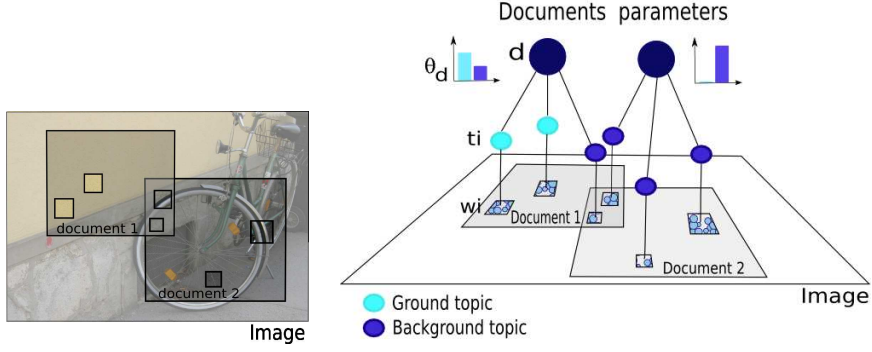


Fig. 3. Example of 2 overlapping documents in an image. 2 of the patches seen as their visual words (w) belong to the object topic (t) and 4 of them belong to the background. The document distribution over topics is represented by the histograms.

ity of such a framework can handle large variations in appearance and shape. However, objects have to cover large parts of the images so they constitute dominant image topics. This is not the case when objects are small as in Figure 2. Furthermore, as no global geometric constraints are used, the proposed model is not well suited for the detection or segmentation tasks.

This is why we propose a new graphical model for representing images and objects, specifically designed to address the problem of object segmentation. Our model, illustrated Figure 3, consists in describing images by a set of overlapping multi-scaled local documents. In this case, even if objects of interest are small, they constitute the main topics of at least a few documents and can therefore be discovered. Each image patch (visual word) belongs to several overlapping documents. The process of assigning labels (figure/ground labels) to each patch is done at the document level, which is a semi-global level. However as documents are overlapping and share image patches, semi-local decisions are propagated all over the image, as MRFs do.

A training stage where some object examples are hand segmented is used to compute the prior for the topic distributions over words used later in the model estimation of unseen images. Consequently, our model is semi-supervised.

4 Image segmentation using the multi-document model

The former model (section 2) is going to be extended to perform class specific object segmentation. The main difference with that model is that each image is now supposed to be a collection of many overlapping documents. This approach can not perform a segmentation task in a fully unsupervised framework; we now assume having distinct training and testing sets. The training set is used to produce the vocabulary, allowing to vector quantize descriptors of test images. Let us now describe the model and explain how to estimate it.

4.1 Description

An image is described by a multitude of different documents ($d \in D$) corresponding to overlapping regions. Documents are chosen to cover uniformly the image. Each document has its own distribution over the topics, denoted θ_d . In contrast, within all documents the probability for topics to generate visual words is the same. Topic distribution over words, denoted ϕ , is sampled from a Dirichlet distribution of hyper-parameter β as before. The model is illustrated Figure 3.

Modeling an image I with this model assumes that it is built according to the following generative process:

- (1) first, the distribution $\theta_d \sim \text{Dir}(\alpha)$ is produced for each document d , where $\text{Dir}(\alpha)$ is a Dirichlet distribution of hyper-parameter α , providing a distribution over the latent *topic* factors,
- (2) for each observation (*i.e.*, a patch associated to a visual word w and a location x) :
 - (a) equiprobably choose a document d from the set of documents containing x . $p(d|x) = 0$ if $x \notin d$ and $p(d|x) = \frac{1}{N}$ if $x \in d$, where N is the total number of documents containing x .
 - (b) draw a topic z from the multinomial distribution of parameter θ_d : $z \sim \text{Mult}(\theta_d)$
 - (c) finally draw a word w conditional on z from the multinomial distribution ϕ , $w \sim \text{Mult}(\phi)$.

The joint probability $p(w, d, z, x)$ is assumed to have the form of the graphical model shown Figure 4. Marginalizing over topics z and documents d determines the conditional probability $p(w|x, \phi, \alpha, \beta, I)$:

$$p(w|x, \phi, \alpha, \beta, I) = \sum_{d \in D} \int_{\theta} \sum_{z \in Z} p(w|z, \phi) p(z|d, \theta) p(d|x) d\theta \quad (3)$$

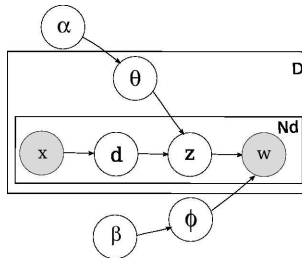


Fig. 4. The graphical model of the segmentation method: an image I is a set of patches, included into overlapping documents d . Each patch is generated by choosing a position x , a document d , a topic t and a visual word w , according to multinomial distributions θ , ϕ .

where w stands for the visual word, x its position, Z the set of latent topics, D the set of documents, I the image and θ and ϕ are the previously mentioned multinomial distributions.

4.2 Model estimation

Our first model (latent mixture vocabulary model) is estimated during a training stage on labeled training images in order to estimate the visual vocabulary and very strong prior on the distribution from topics to words (ϕ). The visual vocabulary is learned once for all during this stage and used to quantify all test image descriptors.

This multi-document model is estimated on test images where we assume that the patch position x and its corresponding visual words w can be directly observed. Hyper-parameters α and β have fixed values like for the previous model. Estimating this model on test images consists in computing the multinomial distributions θ and ϕ with respect to their Dirichlet prior α and β , knowing a set of x and w observed in images. The estimation is done according to the maximum likelihood criterion: we collect N patches from the images and observe the set $(x_1, w_1), \dots, (x_N, w_N)$. We want to compute θ and ϕ maximizing $p((x_1, w_1), \dots, (x_N, w_N) | \theta, \phi, \alpha, \beta)$.

The model given (eq. 3) is again too complicated to be directly estimated, we also used the Gibbs sampling technique for the estimation. During this process we estimate topic assignments (hidden variables of the model) jointly with θ and ϕ . This estimation process is very similar to the one described in section 2.2, indeed it involves the same probability distributions.

We also have to note that for making the estimation possible we only process one image at a time. We typically have thousands of documents per image. Processing all these images simultaneously would be infeasible. As a consequence, documents of different images become independent.

As we said before the training stage is used to acquire strong priors on the distributions from topics to words (ϕ). It gives a good initialization of word instances assignments to topics and guide efficiently the whole process. Nevertheless the ϕ distribution can be adapted to each particular image: the model learns what the topics look like for this specific image.

4.3 From patches to segmentation

At the end of the estimation process, all the patches have a probability of being generated by one of the class topics. These patches correspond to the square sub-window's pixels used to build visual words. To compute the probability for a pixel p to belong to an object class (corresponding to topic z), we accumulate the knowledge on patches \mathcal{P} containing the pixel. This is modeled by a mixture model, where weights (probability of a pixel to have been generated by a patch $p(p|\mathcal{P})$) are function of the distance between the pixel and the center of the patch.

$$p(class(p) = z) \propto \sum_{\mathcal{P}_i \ni p} p(t_i = z)p(p|\mathcal{P}_i) \quad (4)$$

where t_i stands for the topic of patch \mathcal{P}_i .

This can be seen as a summary of all labels provided for the same pixel. In regions where neighboring patches disagree, the confidence will be low; in contrast if neighboring patches agree, the probability for the pixel to belong to the object becomes higher.

5 Experimental results

5.1 Datasets

Experiments have been carried out on five different datasets, illustrated Figure 5.

The first one is a subset of the ETH-80 [18], in which 4 categories have been selected (Apple, Car, Cow and Cup). Each category contains from 10 to 14 objects, from different viewpoints (there are 820 images in total, 205 per category). Despite the fact that these images have not been taken in real conditions (blue background) they are interesting for two reasons. First, the absence of background guarantees that the information used to classify images is not coming from the background but comes from objects themselves (the presence of contextual information can sometimes make the classification task easier). The second interest for using this database is the viewpoints diversity. Building an algorithm able of assigning a top view and a front view of the same object to the same category in an open and interesting issue.

The second database is a Birds dataset [17]. It contains 6 categories and 100

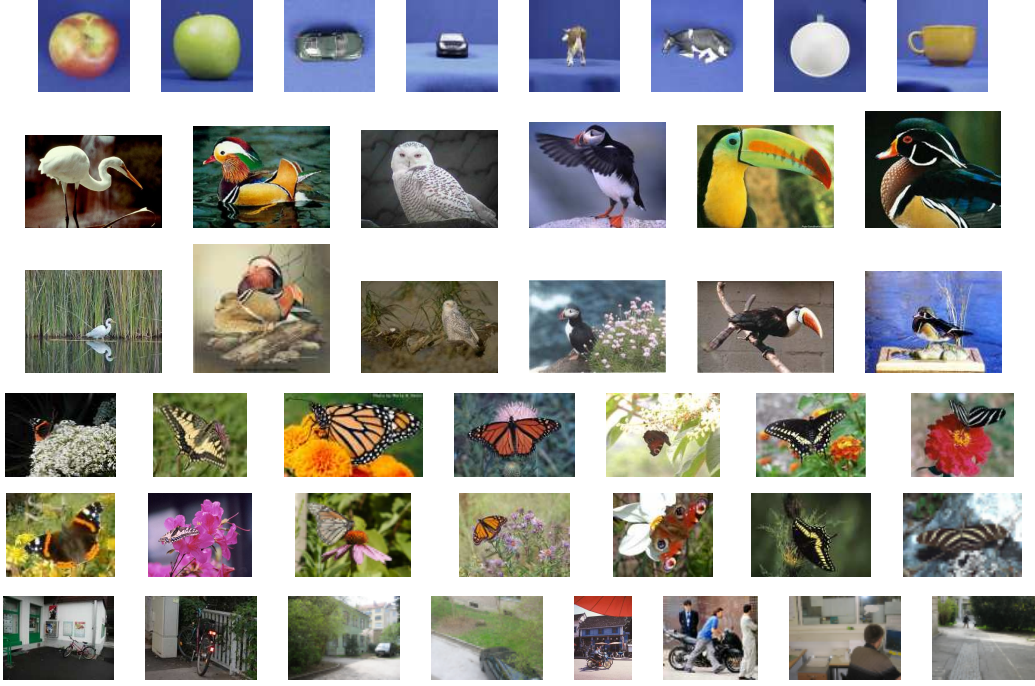


Fig. 5. ETH-80 (first line), Birds dataset (second and third lines), Butterflies dataset (forth and fifth lines) and Pascal dataset (last line): 2 illustrative images per category.

images per category. For classification accuracy evaluation the images were divided into 300 training images and 300 testing images as suggested in [17]. The third one is a Butterflies dataset [16] containing 7 categories and 619 images divided for classification into 182 training images and 437 testing images. The fourth dataset has been released during the Pascal Challenge 2005 [5], containing 684 training images, and 689 testing images. The last database is the bike class of the Graz02 dataset³ whose image examples are shown Figure 2. It contains 365 images with very different bike occurrences.

The large intra-class variability, the scale and viewpoint changes and the highly cluttered backgrounds make these datasets interesting. Finding statistical properties of these images is typically one of the problems addressed by our method.

For all datasets, color information has been discarded and images are considered as grey level images.

³ available at http://www.emt.tugraz.at/pinz/data/GRAZ_02/

5.2 *Experimental settings*

For all of the presented experiments, local descriptors are extracted on a dense grid, at different scales. We do not report here results obtained using interest points detectors which gave worse performances. The setting we used for classification gives approximately 800 patches per image for the ETH dataset, 1500 patches for birds and butterflies and 10 000 patches for the Pascal dataset. For segmentation around 10 000 patches per image were extracted. This sampling is dense enough for all pixels to belong to several patches. Each patch is represented by a 128 dimensional SIFT descriptor [19].

We assumed that the Dirichlet priors are symmetric, α and β having a fixed scalar value as suggested by [10].

We observed that the Gibbs sampler “converges” after less than 50 iterations, which is the number used for these experiments. It takes about 12 hours to learn the latent mixture vocabulary model. If the vocabulary has already been learned in a training step, it takes few minutes to segment an image. It is also important to note that, in order to reduce the amount of memory required to store visual descriptors, we vector quantized them.

All reported classification accuracies are obtained combining 1 vs 1 linear SVM classifiers. Multi-class performances are considered except for the Pascal dataset which involves binary classification. We report both means and variances of 5 runs with different random initializations. Except when specified, we used a visual vocabulary of 1000 words.

5.3 *Image classification experiments*

The section assesses the superiority of the vocabularies built by the proposed method. Experiments are divided into two separate problems: image categorization based on latent topics and image categorization using visual features in a bag-of-features framework. The same model is used in both cases, but different information is given to the classifier.

Two baseline methods have been implemented for comparison purposes: the standard bag-of-features approach (using k-means to build the visual vocabulary), and the standard LDA model (also using k-means to build the vocabulary).

5.3.1 *Topic based image categorization*

Ideally, latent based methods can be completely unsupervised as it has been shown by [23]. The number of topics can be fixed as being the number of actual categories and each category is then represented by only one topic.

However we argue that classes are highly semantic concepts and rely more on human knowledge than visual characterization. Indeed, we observed during these experiments that except in very simple cases, estimated topics rarely coincide with true classes. More precisely, there are many local minima making the outcome of the process very depending on initialization; topics match with categories for only a few of these modes. One solution can be to use topics in a more supervised framework, as described in section 2.3. In this case, class labels were used to reduce the number of parameters of the model, making its estimation a more convex problem. Then we can use a simple Bayesian Classifier assigning the label of the most probable topic (TOPIC-BAYES) or use a classifier considering topic distributions as feature vectors. The classifier is trained on images which were labeled for learning. We denote this classification scheme TOPIC-SVM.

Using these two topic-based strategies, topics produced by our model (denoted LDA-VOC) were compared to a baseline LDA model which does not learn the vocabulary (denoted STD-LDA, for standard LDA).

Table 1 summarizes these experiments on the two first datasets. Each line corresponds to a different amount of supervision, from 0 labeled images (fully unsupervised case, which is not applicable with TOPIC-SVM which requires at least 1 labeled image per class) up to a larger number. Without any supervision the variance is very high in best cases (ETH-80) while in worst cases (birds datasets) the classification is not possible as topics are not related to categories at all. The supervision helps the system to produce better and more stable (low variance) results and should not be considered as optional.

It is important to note that with both datasets and under all of the different settings LDA-VOC performs much better than STD-LDA. We also note that TOPIC-BAYES and TOPIC-SVM performs equally.

Results on the ETH-80 dataset are impressive; despite the large number of viewpoints, giving only 2 labeled images per category is enough for grouping all of the viewpoints of the same category. The Birds dataset is much harder and even with a large amount of supervision the performances are rather low. It gave us the feeling that topics could not be the best information for classifying images, especially if only a few topics are considered and if images present a highly cluttered background.

5.3.2 Bag-of-features image classification

In these experiments we estimate the model exactly as it has been done in the previous section. However, instead of classifying images using their topic distributions we trained a bag-of-features classifier using the vocabulary produced by our model. We focused our experiments on comparing the standard bag-of-features approach using k-means to quantize the feature space and a linear SVM classifier, denoted KMEANS-BOF, with the bag-of-features which uses the vocabulary produced by our model, denoted LDA-VOC-BOF (see section 2.3) and the same SVM classifier.

For this purpose we split the datasets in two parts (training and testing). The training part, which is labeled, is the supervised part in the model learning and is used to train the classifiers. For the ETH dataset, which is a much easier dataset, a small number of training images is enough to reach 100% accuracy with any version of the bag-of-features approach so we discarded this dataset for these experiments. We report in Table 2 the mean of classification results obtained for different vocabulary sizes.

From these results we can draw several remarks. First, we note that the vocabulary given by our model is better: the overall classification rate can be

ETH-80	TOPIC-BAYES				TOPIC-SVM			
nb labeled	LDA-VOC		STD-LDA		LDA-VOC		STD-LDA	
img	Av	var	Av	var	Av	var	Av	var
0	88.92%	12.43	-	-				
8	96.42%	1.53	94.62%	0.05	96.8%	1.12	94.6%	0.18
176	98.73%	0.08	97.16%	0.03	98.72%	0.25	97.19%	0.15

BIRDS	TOPIC-BAYES				TOPIC-SVM			
nb labeled	LDA-VOC		STD-LDA		LDA-VOC		STD-LDA	
img	Av	var	Av	var	Av	var	Av	var
0	-	-	-	-				
66	44.01%	0.21	-	-	43.6 %	0.26	39.1%	0.46
198	55.97%	0.2	50.3%	1.01	55.6%	0.22	50.3%	1.02
300	60.68%	0.72	54.5%	0.6	60.67%	0.75	54.4%	0.75

Table 1

TOPIC-BAYES and TOPIC-SVM results for the ETH-80 (top) and Birds (bottom) datasets. Each line represents a different level of supervision (labeled images). We report average performance as well as variance. “-” means that topics can not be assigned to classes for at least one of the run.

nb of words	BUTTERFLIES					BIRDS				
	LDA-VOC-BOF		KMEANS-BOF			LDA-VOC-BOF		KMEANS-BOF		
200	76.2 %		67.89 %			74.6 %		65.33 %		
500	83.83%		78.57 %			85.1 %		76.58 %		
1000	88.56%		84.65 %			89.0 %		83.33 %		
2000	90.38%		85.77 %			90.9 %		86.17 %		

nb of words	Pascal Challenge									
	LDA-VOC-BOF					KMEANS-BOF				
class	Cl 1	Cl 2	Cl 3	Cl 4	mean	Cl 1	Cl 2	Cl 3	Cl 4	mean
200	87.8	90.2	93.4	86.9	89.6	86.8	88.4	89.6	83.3	87.0
500	89.6	91.5	95.3	90.5	91.7	86.0	91.3	96.6	82.1	89.0
1000	89.5	92.7	97.0	91.2	92.6	89.7	92.7	96.6	89.9	92.3

Table 2

Comparing the vocabulary produced by our model (LDA-VOC-BOF) with a vocabulary obtained by a k-means quantization of the feature space (KMEANS-BOF).

increased by nearly 10%.

Second, using bag-of-features instead of topic based classification leads to better results (a gain of more than 30% for Birds), which can be explained by the coarseness of the model. These experiments also confirm our feeling that, in some situations, classifying images using words statistics can be better than using topic distributions.

Third, the overall performance of our system is very similar to the best results reported on the Birds dataset [17], although we do not use any geometric information. The total classification rate for butterflies is also comparable to the best known results [16] and our method is much better in terms of average class accuracy. Confusion matrices for runs with a 2000 words vocabulary can be found Table 3.

Finally, the results obtained for the pascal dataset using the bounding boxes as supervision are displayed on the second part of Table 2. The improvement seems lower for two-classes problem than in the multiclass framework. We still observe a difference between the two methods, which is more significant for small vocabularies.

We also tried to increase the number of topics, in a range from the number of category to larger numbers and we noticed that the behavior of the system moved from the behavior of the topic based classifier to the behavior of the

bag-of-features classifier.

5.4 Analyzing the vocabulary

Our main motivation for learning the vocabulary simultaneously with other parameters was to produce visual words that should be more adapted to visual categories. We used a probabilistic criterion to evaluate this adaptation.

We computed $p(C|w)$, the probability of class C given that word w is detected. We histogrammed these values for each class, for all visual words. The top-left part of Figure 6 shows the histogram corresponding to the first category of the birds dataset (similar results have been obtained with other categories). We can see that our model has been able to find more than 20 words for which $p(C|w) > 0.9$, being therefore class specific words whereas k-means gives only 1 discriminative visual word.

	C_1	C_2	C_3	C_4	C_5	C_6	
C_1	43	3	1	1	2	0	86%
C_2	3	45	0	1	1	0	90%
C_3	1	0	49	0	0	0	98%
C_4	0	0	1	49	0	0	98%
C_5	1	1	0	1	47	0	94%
C_6	0	3	0	1	0	46	92%
						Av	93%

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	
C_1	79	0	2	0	1	2	1	92.9%
C_2	0	16	0	0	0	0	0	100%
C_3	0	2	52	2	0	1	0	91.2%
C_4	2	0	0	41	5	0	0	85.4%
C_5	1	0	1	9	47	0	0	81%
C_6	3	0	2	0	0	103	0	95.4%
C_7	3	0	0	4	0	0	58	89.2%
							Av	90.61%

Table 3

Confusion matrix of the best run on the birds and on the butterflies datasets. Number of images and percentages are presented.

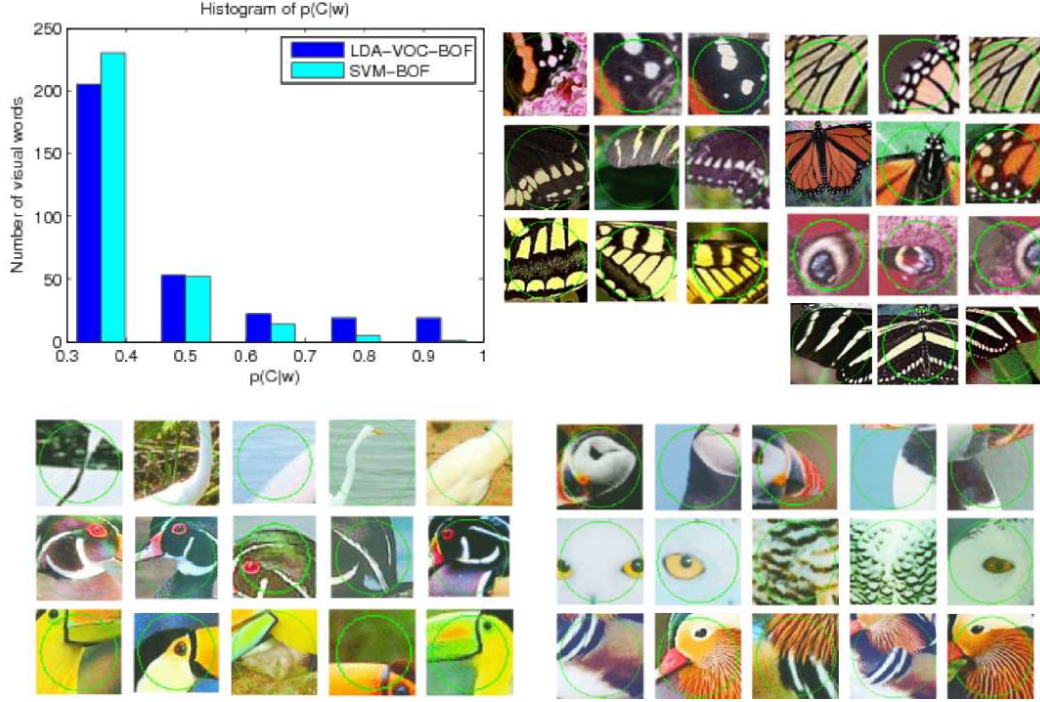


Fig. 6. Top row: (left) density of class probabilities conditional to the words, (right) 3 highly probable words for each of the butterflies class. Bottom row: best 5 visual words per topic for the Birds.

As an illustration, the rest of Figure 6 shows 5 of the most discriminative words per topic for the birds dataset, and also examples of words generated by our model for the butterfly dataset. We can see the vocabulary ability to catch useful class specific information.

5.5 Object level segmentation experiments

For the segmentation experiments we consider only two topics, one for the foreground (object we are interested in) and one for the background (everything else). We keep the multiclass vocabulary but we re-estimate the ϕ distribution according to these two topics.

We will first consider some qualitative results on the database previously used for classification and then quantitative results on the Graz dataset where a pixel level evaluation is possible.

5.5.1 Qualitative results on the Birds and Butterflies dataset

These experiments are performed on the Birds and Butterflies datasets. The images are represented with a 2000 words vocabulary build by our latent mix-

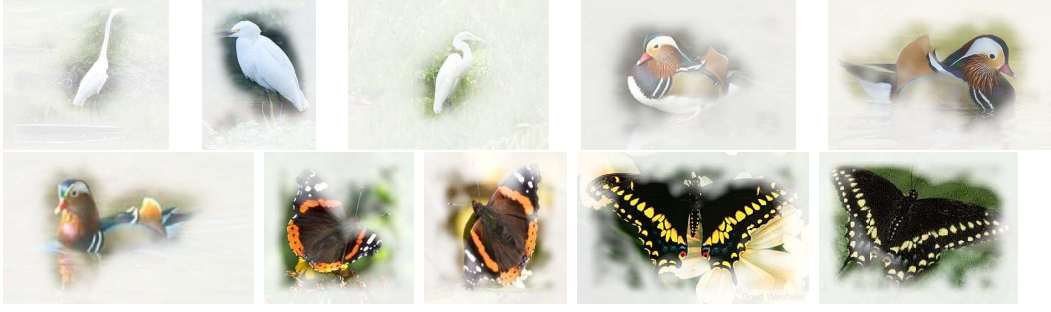


Fig. 7. Examples of probability maps obtained by our method. The image is opaque where pixels have a low probability of being foreground pixels. It is transparent otherwise.

ture vocabulary model. We used bounding boxes to indicate approximately in which part of the image the object is located and then re-estimated foreground and background topics.

The method estimates for each test image patch a probability of belonging to the object class. By summarizing over all patches we can estimate a probability map at the pixel level. Figure 7 shows several examples of pixel level probability maps obtained by our segmentation method. These probability maps are given as transparency layers so the more probably the pixel belongs to an object, the more transparent the layer is. We can see that despite important variations in appearance and shape, the probability maps allow to give a location and an approximate shape of our class instances.

However, on several images, segmentation is really poor or focus only on the most discriminant part of the animal, the rest staying quite undecided. It can be explained by 2 reasons. First the supervision is weak, bounding boxes containing the class instances might not be informative enough to allow accurate estimation of the topics. Second, the animals are often observed on the same sort of background (their living environment) and this background might therefore be learned as a part of the class.

5.5.2 Quantitative results on the Graz dataset

We evaluate our method by comparing segmentation it produces with hand ground truth segmentation. For the Graz2 dataset, the ground truth is available for 300 bike images. It is given in terms of pixel segmentation masks. These masks will be used to evaluate the quality of our segmentation. We will compare them to the probability maps produced by our method and compute an accuracy score (see Figure 8 for some examples).

We have shown section 4.3 that our algorithm computes the probability for

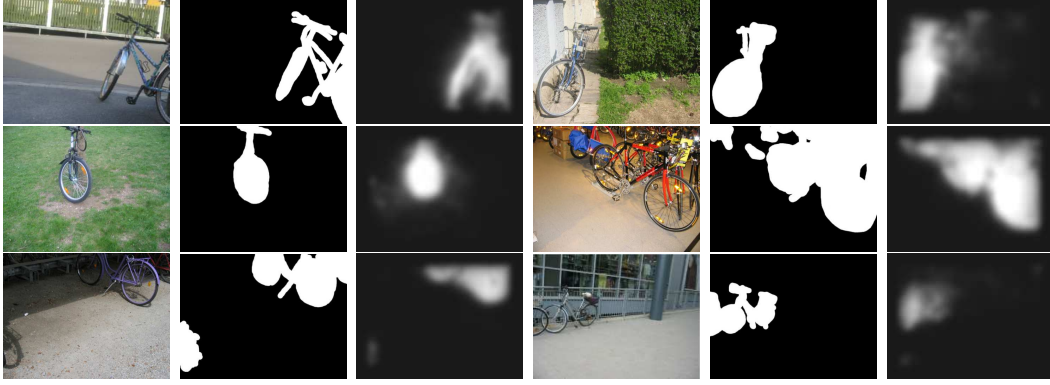


Fig. 8. Several bike images of the Graz02 dataset (left), the ground truth (middle) and the probability map (right) generated by our method.

each image’s pixel to belong to an object of a given category (summarized by the probability map). On the other hand, we know ground truth pixels labels, given by the provided segmentation masks. It is therefore natural to evaluate the performance by computing a ROC curve for each image. The ROC curve represents the true positive rate (TPR) against the false positive rate (FPR)⁴, *i.e.*, the rate of correct classification for the category of interest against the rate of misclassified object pixels. The true positive rate at equal error rate (EER) is the true positive rate at the curve point where $TPR = 1 - FPR$.

First we would like to see, in which measure, using multiple overlapping documents per image is better than a more trivial model. The experiments are conducted with a scenario where all training patches labels are known using the ground-truth segmentation masks of the training images. For comparison purposes we then developed 2 baseline methods:

- (1) *patch based method*: on training images the segmentation masks are used to fix the topic assignments and then estimate the probability for each topic to generate a particular word. We use $p(t|w) = \frac{p(w|t)p(t)}{p(w)}$ to compute the probability for an observed word to belong to the foreground. Pixel level topic estimation is done in the same way as described in subsection 4.3.
- (2) *single document method*: the whole image is considered to be a single document, which is the traditional way of doing it. Except for the mixture of documents, the rest of the method is the same.

We compared our method with the 2 baseline methods. On Figure 9, for different images (upper-right part), the upper-left part shows the ROC curves obtained for the 3 methods. We also show the probability maps for the pro-

⁴ $TPR = \frac{TP}{P}$ and $FPR = \frac{FP}{N}$ where TP , FP , F and P are respectively the numbers of true positive, false positive, positive and negative

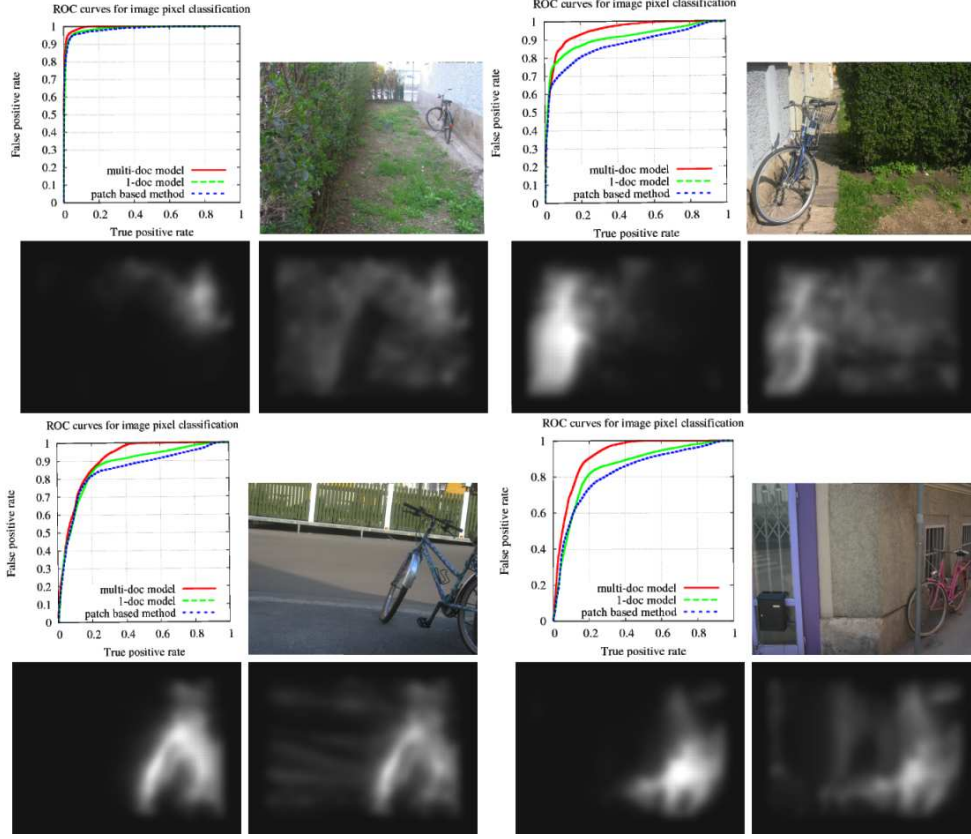


Fig. 9. Comparison of our method with 2 baseline methods. For each original image (upper-right), the ROC curve for the 3 methods (upper-left), the probability map obtained by our method (bottom-left) and the one obtained with the second baseline method: the single document one (bottom-right)

posed methods (bottom-left part) and for one of the 2 baseline methods: the one with one document (bottom-right part). The other baseline map is similar or slightly worse. These results typically illustrate the gain obtained using multiple overlapping documents in the image.

It is interesting to analyze the amount of supervision necessary to have good results. To that aim, We did several experiment with different amounts of supervision:

- (1) segmentation masks which allow to mark precisely if image patches belongs to the object
- (2) bounding boxes which give rectangles containing the objects
- (3) image label: the only information we have is that the object is contained somewhere in the image, without any location information.

These 3 different supervisions were combined in different settings summarized in Table 4. The question is how much the quality of the results depends on supervision.

sup1	segmentation masks for all images	0.79760
sup2	half segmentation masks, half bounding boxes	0.79173
sup3	25% segmentation masks, 75% bounding boxes	0.78711
sup4	bounding boxes for all images	0.76078
sup5	half with bounding boxes, half with no location information	0.72867
sup6	25% with bounding boxes, 75% with no location information	0.63948

Table 4

For each different supervision framework, its description and the mean EER obtained on the different ROC curves of pixel level classification

For each of these supervision frameworks we produced probability maps for test images. Each probability map is used to compute a ROC curve and associated to an EER. The EER average produced on all test images is displayed in the last column of Table 4 for the different supervision settings.

As expected, we observed that more supervision gives more precise segmentation. The table shows that few masks are enough to insure a stable estimation of the image topic and then produce satisfying segmentation. Even bounding boxes give reasonable results even if the loss of accuracy is noticeable. As an illustration, Figure 10 shows the masks for the different amount of supervision described in Table 4.

It is interesting to analyze the results obtained on all test images. We computed the average ROC curve per supervision method with error bars representing the standard deviation and show it Figure 11 for 3 different settings (sup1, sup4, sup5). The standard deviation could seem to be important at a first glance. This can be easily explained by two different factors. First, we measure performance with ROC curves; as the number of foreground pixels is very small (often less than 10% of the total number of pixels) compare to the number of background pixels, even if only a small fraction of the foreground pixels are misclassified it impacts the ROC curve significantly. The second factor is the variability and the difficulty of the Graz02 images: some objects are barely detectable, even by humans. At the equal error rate, the recall is nearly 80% for the strongest supervision framework and falls to 76% for bounding boxes.

At last, we show Figure 12 binary segmentation masks obtained by thresholding the probability map. We can see the accuracy of the method, considering that no strong cues (color, texture, shape, etc.) are here to give evidence for pixels to belong to the objects of interest.

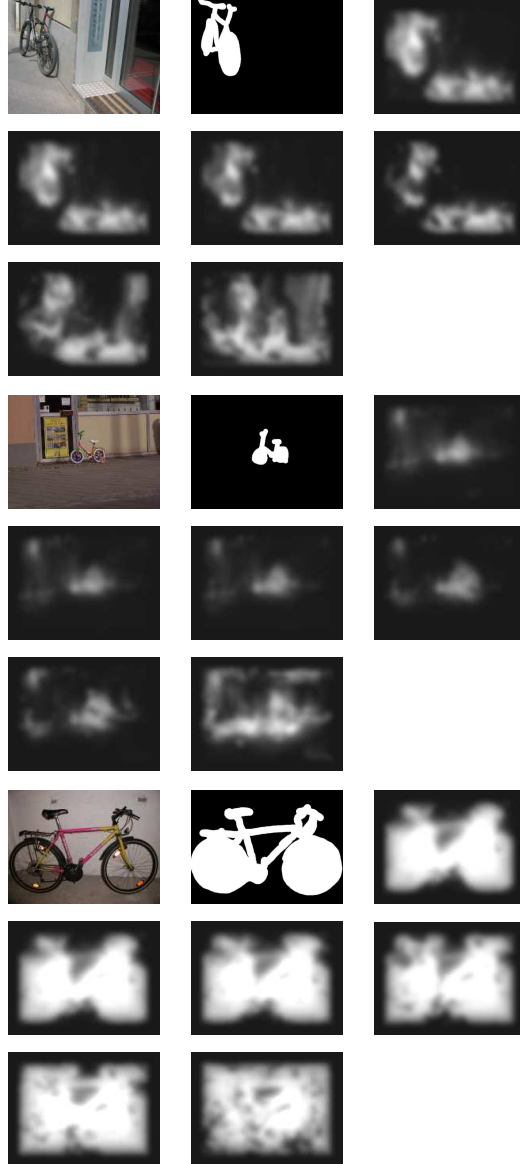
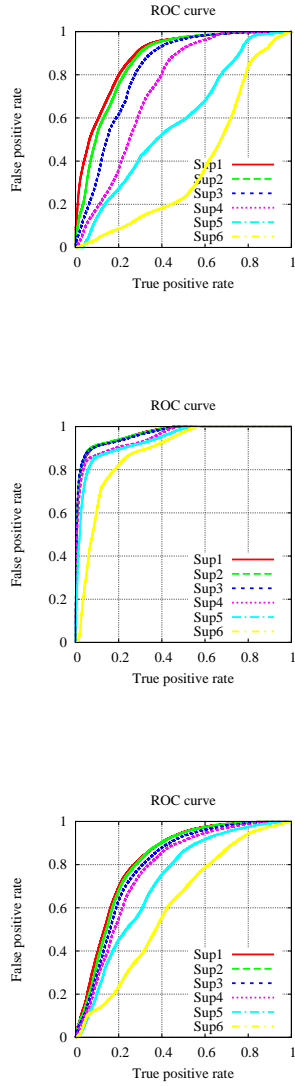


Fig. 10. probability maps for different levels of supervision, and the corresponding ROC curves on some typical bike images. The images are first the original image, then the ground-truth and finally the probability maps for the settings from 1 to 6.

6 Conclusions

In this paper we presented a new framework for creating visual vocabularies, in the context of object categorization and object segmentation problems. The core of this framework is an object model embedding visual words as a component of the learning process.

It was experimentally shown on different datasets that this model outperforms methods for which the vocabulary is built separately. The number of words

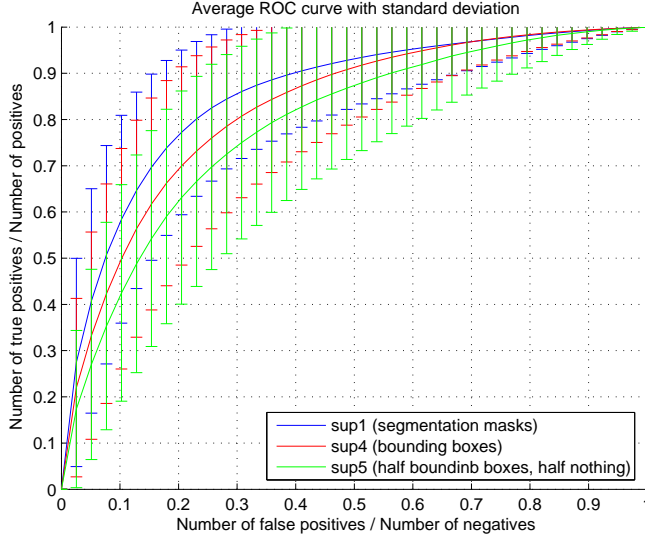


Fig. 11. Average ROC curve obtained on all training images, with the standard deviation (error bars).

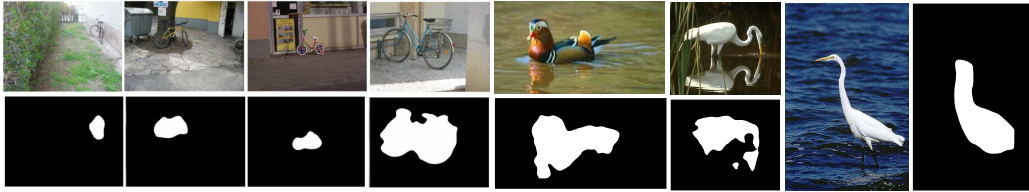


Fig. 12. Examples binary segmentation obtained with our method on the bike category of the Graz 02 dataset and on the Birds dataset.

used for getting good performances is lower than standard bag-of-features approaches. It is due to the model ability to quantize the descriptor space in a smarter way than a standard clustering method. The words are more adapted to the task and more focused on class discriminative information.

As all observations are simultaneously used by our model, learning parameters is much more time consuming than standard LDA or basic clustering methods.

Another conclusion is that the bag-of-features approach outperforms the topic based classifiers, especially if a large amount of training data is available.

We have also proposed a new method for learning to segment objects in images which is an extension of the method used for classification. It considers images as being made of multiple overlapping regions, treated as distinct documents. Used in a semi-supervised framework, it can achieve a remarkably high precision even with difficult images.

However, further improvements could possibly make the performances even better. One of these improvements would be to embed within the model local

shape information, making the detection of object boundaries more accurate.

References

- [1] D. Blei and M. Jordan. Modeling annotated data. In *SIGIR*, 2003.
- [2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *NIPS*, 2002.
- [3] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In *Proc. of CVPR Workshops*, 2004.
- [4] G. Csurka, C. Dance, L. Fan, J. Williamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV workshop on Statistical Learning in ComputerVision*, pages 59–74, 2004.
- [5] M. Everingham, A. Zisserman, C. K. I. Williams, L. J. Van Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkó, S. Duffner, J. Eichhorn, J. D. R. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A. J. Storkey, S. Szedmák, B. Triggs, I. Ulusoy, V. Viitaniemi, and J. Zhang. The 2005 pascal visual object classes challenge. In *MLCW*, pages 117–176, 2005.
- [6] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.
- [7] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google’s image search. In *ICCV*, volume 101, pages 5228–5235, 2005.
- [8] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Qian Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: the qbic system. *Computer*, 28(9):23–32, 1995.
- [9] M. Fritz, B. Leibe, B. Caputo, and B. Schiele. Integrating representative and discriminative models for object category detection. In *ICCV*, 2005.
- [10] T. Griffiths and M. Steyvers. Finding scientific topics. In *PNAS*, 2004.
- [11] R.M. Haralick and L.G. Shapiro. Image segmentation techniques. *CVGIP*, 29:100–132, 1985.
- [12] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, 2001.
- [13] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, 2005.
- [14] M. P. Kumar, P. H. S. Torr, and A. Zisserman. OBJ CUT. In *CVPR*, 2005.
- [15] S. Kumar and M. Hebert. Discriminative random fields. *IJCV*, 68(2):179–201, June 2006.

- [16] S. Lazebnik, C. Schmid, and J. Ponce. Semi-local affine parts for object recognition. In *BMVC*, volume 2, pages 779–788, 2004.
- [17] S. Lazebnik, C. Schmid, and J. Ponce. A maximum entropy framework for part-based texture and object recognition. In *ICCV*, 2005.
- [18] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC*, 2003.
- [19] D. J. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
- [20] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *ECCV*, 2006.
- [21] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *NIPS*, pages 1097–1104, 2004.
- [22] P. Quelhas, F. Monay, J. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. In *ICCV*, 2005.
- [23] J. Sivic, B. Russell, A. Efros, A. Zisserman, and B. Freeman. Discovering objects and their location in images. In *ICCV*, 2005.
- [24] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, volume 2, pages 1470–1477, October 2003.
- [25] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Describing visual scenes using transformed dirichlet processes. In *NIPS*. 2006.
- [26] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.
- [27] J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *CVPR*, pages 37–44, 2006.