Latent Mixture Vocabularies for Object Categorization

Diane Larlus and Frédéric Jurie
LEAR Group, INPG - CNRS, INRIA Rhônes Alpes, France
{diane.larlus, frederic.jurie}@inrialpes.fr

Abstract

The *visual vocabulary* is an intermediate level representation which has been proven 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. We experimentally show that the proposed model outperforms approaches not learning such an adapted visual vocabulary.

1 Introduction

Object categorization is an important task in computer vision, which has received a lot of attention over the last three years [3, 4, 6, 9, 10, 14, 15, 16, 18]. This problem is challenging because of pose and illumination changes, scale variations, occlusions and intra-class variability, which potentially make two images of the same class very different.

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.

Very efficient statistical models have been used to address this problem, models often inspired by text analysis. After building a visual vocabulary, images can be processed as sets of visual words therefore 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 [3] and [16], and later extended by many other authors like [10, 18]. Images are simply modeled by measuring frequencies of unordered sets of visual words, encoded as histograms.

The impressive bag-of-features strategy inspired more complex models, like the probabilist Latent Semantic Analysis (pLSA) [8] or its Bayesian form, the Latent Dirichlet Allocation (LDA) [2]. These models have recently been applied to object categorization [4, 5, 14, 15, 17]. 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 [12]), and vector quantized.

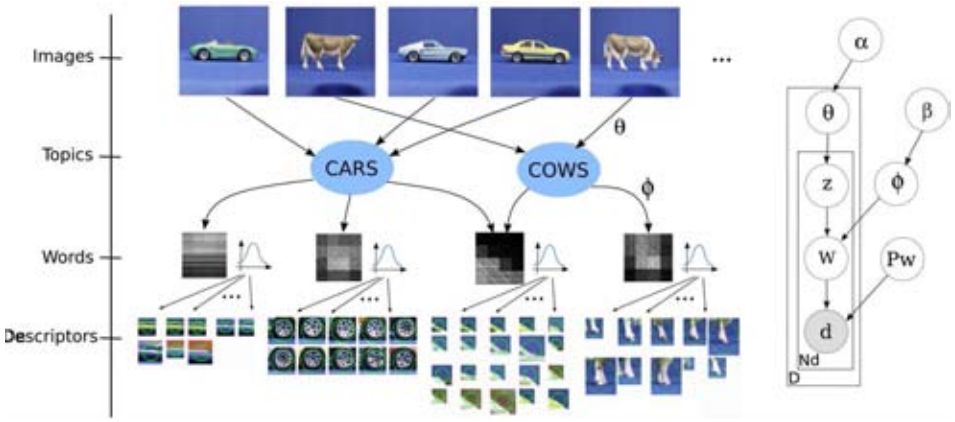

Figure 1: Overview of the method, and the corresponding graphical model representation.

Several combination of visual descriptors and clusterers 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 [3, 16]. Agglomerative techniques [11] or mean-shift based approaches [9] 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. On the other hand, 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 [18], 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 very recently by Perronnin *et al.*. [13]. They address this issues by combining universal vocabularies with specific vocabularies. The universal vocabulary describes the visual content of all the considered classes while the class specific vocabularies are obtained by adapting the universal vocabulary to a class using specific data. This combination of universal and specific approach 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. Therefore, if two classes are visually close, there is no guarantee to obtain distinctive visual words.

The approach proposed in this article tries to go one step further in the aim of producing distinctive visual vocabularies. Inspired by [14, 15], 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 of *topics*, topics as distributions of *words* and words as Gaussian

mixtures of visual descriptors (see Figure 1 for an illustration of the model).

By imposing Dirichlet priors on topic and word distributions the model is incited to produce a few but 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 article is as follows. Section 2 presents the proposed model and the way to estimate it. Section 3 explains how to use the model parameters to classify images. Our model can be used with different classification strategies, which are described in this section. In Section 4 we present our experiments, and finally, the conclusions are drawn.

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 images[1]. In this article, visual descriptors are SIFT vectors in a 128-dimensional space, but other descriptors could be used. Position and scale of these descriptors in the image are discarded.

We use a simplified form of Gaussian-Multinomial LDA (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 combination of T possible topics which themselves produce N visual words, and words are Gaussian distributions over the SIFT descriptor space. Topic distributions over words (ϕ) are sampled from a Dirichlet distribution of parameter β.

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 of hyper-parameter α, providing a distribution over the latent *topic* factors,
2. For each of the image descriptor d_i,
   (a) sample a topic z_i from the multinomial distribution of parameter θ : $z_i \sim Mult(\theta)$.
   (b) sample a visual word w_i conditional on z_i from the multinomial distribution of parameter ϕ_{z_i}, $w_i \sim Mult(\phi_{z_i})$
   (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 of parameter P_{w_i}.

The resulting distribution on visual descriptors in image I is given as follows:

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

[1] In practice, according to [9] and [18] we pick patches on a regular grid at multiple scales.

Compared to [4, 5, 14, 15] 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

Learning the model consists in a likelihood maximization and is done by estimating the optimal parameters α, β, ϕ et θ for a given set of images. Hyperparameters α and β play an important role as they allow, by using particular values, to control how topics and visual words distribution can be sparse and therefore specialized. This is why, according to [7] we prefer not to estimate them and use fixed Dirichlet priors.

Since the integral in equation (1) makes the direct optimization of the likelihood intractable, we estimate variables of interest by an approximate iterative technique called Gibbs Sampling. Our estimation method is very similar to [7][2] which presents an efficient algorithm for LDA estimation.

The parameters of the model can be estimated without any need for supervision, i.e. using unlabeled training images. It is expected that if we marginalize $P(\theta,\phi,P,d_i)$ over ϕ,P and d_i, $P(\theta)$ 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 occupied only a small fraction of the image. In this case, the Gibbs sampler can get stuck in one of these "bad" modes, depending on initialization. We observed that these unwanted modes can be suppressed by adding supervision, assuming topics are known[3] for a few training images. In all cases, we experimentally observed that this kind of supervision leads to more accurate estimation and improves performance a lot.

3 Classifying images

Once the model is learned, image classification can be achieved in different ways.

Topic based maximum Likelihood A natural decision rule is the Maximum Likelihood rule (ML)[4]. The most straightforward way to implement this rule is to set the number of topic equals 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 θ_i in image I, we have $p(C_i|I) = p(\theta_i|I)$. In practice, instead of using equation (1) and integrating over θ, we use the output of the Gibbs sampler and approximate the integral by a sum of discrete values. We denote this rule the TOPIC-BAYES classifier.

Topic based SVM classifier However, if we want more topics than object classes the TOPIC-BAYES rule cannot be applied anymore. We adapted the more general classification scheme proposed by [14] to our model. This scheme consists in training a classifier on the latent variables associated with each image. This cannot be directly used

[2] justification and implementation details of topic estimation process can be found in this reference

[3] derived from class labels

[4] as prior on classes are generally not available, we do not consider here the Maximum A Posteriori criterion

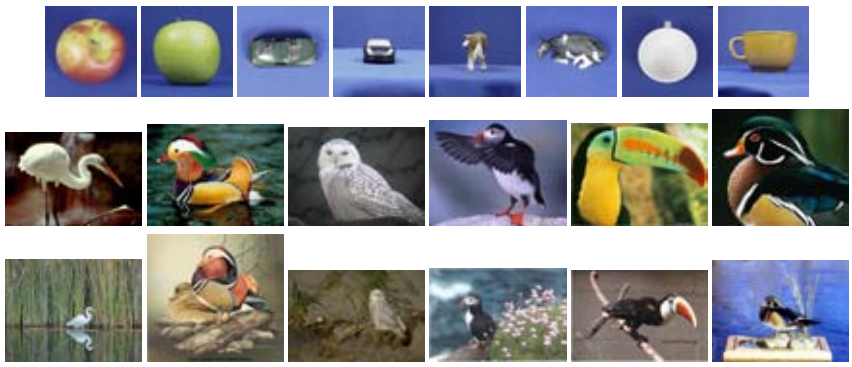

Figure 2: ETH-80 (top) and birds dataset (bottom): 2 illustrative images per category.

with our Gaussian-LDA model which does not explicitly estimates numerical values for latent variables but probability densities. We experimentally observed that the Markov chain generated by the Gibbs sampler tends to converge quickly towards sharp and stable modes. We assign the values corresponding to these modes to each image, and we train an SVM classifier on these values. We call this classifier the TOPIC-SVM classifier.

Bag-of-features based SVM classifier Instead of classifying images from their topic distribution, image can also be classified according to their visual words statistics, as it was done in the original bag-of-features approach. 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.

4 Experimental results

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 parameters are used by the classifier.

As a secondary issue, we also compare the performance of the topic based classification scheme versus the bag-of-features one, for different amounts of supervision.

Two baseline methods have also 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).

4.1 Datasets

Experiments have been carried out on two datasets, illustrated Figure 2. The first one is a subset of the ETH-80 [11], 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 and therefore actually 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 bird dataset [10]. It contains 6 categories and 100 images per category. The large intraclass variability, the scale and viewpoint changes and the highly cluttered backgrounds make this dataset interesting. Finding statistical properties of these images is typically one of the problems addressed by our method.

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

4.2 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 give approximately 800 patches per image for the ETH dataset and 1500 patches for the birds dataset. Each patch is represented by a 128 dimensional SIFT descriptor [12].

We assumed that the Dirichlet priors are symmetric, α and β having a fixed scalar value. This prior knowledge on multinomial distributions controls the mixing of the multinomial weights. Using low hyper-parameters encourages the distributions to be sparser. Images will then more probably choose having a small number of topics, and topics a few number of words. We used for these experiments $\alpha_i = \beta_i = 0.5, \forall i \in \{1,..,T\}$.

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 process each of our databases. 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 results are multiclass performances obtained by combining 1 vs. 1 SVM classifiers. We report both means and variances of 5 runs with different random initializations. Except when specified, we used a visual vocabulary of 1000 words.

4.3 Topic based image categorization

Ideally, latent based methods can completely be unsupervised as it has been shown by [15]. 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 a highly semantic concept and rely more in human knowledge than visual characteristics. 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 depended 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 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 vectors as feature vectors. The classifier is trained on images which were labeled for learning. We denote this classification scheme TOPIC-SVM.

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

Figure 3: 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.

Figure 4: As an illustration, the most discriminative patches used for classification are shown for one of the birds class on few images.

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).

Figure 3 summarizes these experiments on the two datasets. Each line corresponds to a different amount of supervision, from 0 labeled images (fully unsupervised case[5]) 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 bird 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.

4.4 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 focussed 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 use the vocabulary produced by our model, denoted LDA-VOC-BOF (see section 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

[5] not applicable with TOPIC-SVM which requires at least 1 labeled images per class

	ETH-80		BIRDS	
	LDA-VOC-BOF	KMEANS-BOF	LDA-VOC-BOF	KMEANS-BOF
200 words	87.7 %	87.5 %	74.6 %	65.33 %
500 words	87.4 %	86.9 %	85.1 %	76.58 %
1000 words	84.9 %	85.1 %	89.0 %	83.33 %
2000 words			90.9 %	86.17 %

Table 1: 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).

the classifiers. For the birds dataset we defined the training and testing sets as specified by [10] (300 images per set). For the ETH dataset, which is a much easier dataset, the training set includes only 12 images, the remaining ones being kept for testing. We report in Table 1 the mean of classification results obtained for different vocabulary sizes.

From these results we can draw three remarks. First, we note that the vocabulary given by our model is much better: the overall classification rate can be 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. Third, the overall performance of our system is very similar to the best results reported on the Birds dataset [10], although we do not use any geometric information.

These experiments also confirm our feeling that, in some situations, classifying images using words statistics can be better than using topic distributions. We wished to go further and tried to outline the limitations of these methods for different number of training images. Our feelings was that the bag-of-feature approach can reach higher performances but requires more training images, because of the higher dimensionality of the vector space. Our experiments, illustrated by Figure 5, confirmed these feelings. In order to obtain these results, we used a model learned with 12 labeled images, considered both at topic and word levels. We added a variable number of labeled images to train the classifier : when enough training images are available, the bag-of-features performs better than the topic based classifier.

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

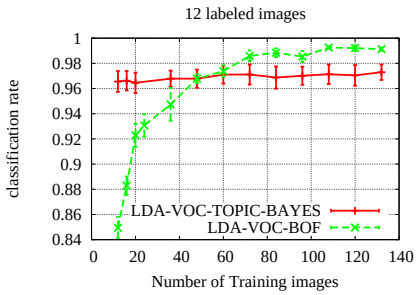

Figure 5: ETH-80 : comparing topic based classification (TOPIC-BAYES) with bag-of-features (LDA-VOC-BOF), as a function of the number of training images. Both representation come from the same model.

	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%

Figure 6: Confusion matrix of the best run on the birds dataset. Number of images and percentages are presented.

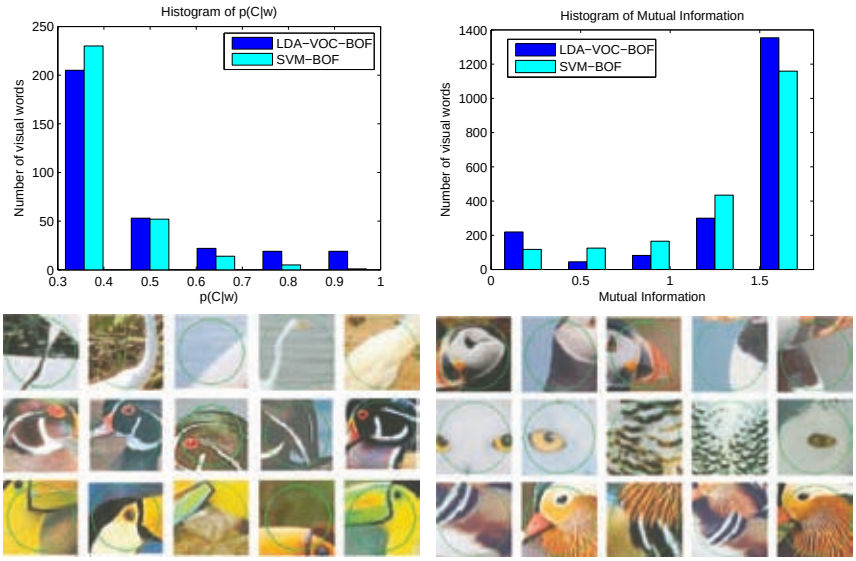

Figure 7: Top row: (left) density of class probabilities conditional to the words, (right) density of information between words and classes. Bottom row: best 5 visual words per topic.

4.5 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. Two different criteria have been taken into account to evaluate this adaptation.

First, we computed $p(C|w)$ which is the probability of having the class C when the word w is detected. We histogrammed these values for each class, for all visual words. The top-left part of Figure 7 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 a k-means quantization gives only 1 of these discriminative visual words.

We also computed the mutual information between classes and visual words and show the corresponding histogram on the top-right part of Figure 7. Words with low entropy, which are those well correlated with classes, are almost twice as numerous as those obtained with k-means.

As an illustration the bottom part of Figure 7 shows the 5 most discriminative words per topic. We can see the vocabulary ability to catch useful class specific information.

5 Conclusion

In this paper we presented a new framework for visual vocabulary creation in object categorization context. The core of this framework is an object model embedding visual words as a component of the learning process.

It was experimentally shown on two different datasets that this model outperforms methods for which the vocabulary is built separately. The number of words used for getting good performances is lower to 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 every observation is explained by our model, its estimation is much more time consuming than a standard LDA or a basic clustering method.

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

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] G. Csurka, C. Dance, L. Fan, J. Williamowski, and C. Bray. Visual categorizationwith bags of keypoints. In *ECCV workshop on Statistical Learning in ComputerVision*, pages 59–74, 2004.

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

[5] 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.

[6] M. Fritz, B. Leibe, B. Caputo, and B. Schiele. Integrating representative and discriminative models for object category detection. In *ICCV*, 2005.

[7] T. Griffiths and M. Steyvers. Finding scientific topics. In *Proc. of the National Academy of Sciences*, 2004.

[8] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196, 2001.

[9] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, 2005.

[10] S. Lazebnik, C. Schmid, and J. Ponce. A maximum entropy framework for part-based texture and object recognition. In *ICCV*, 2005.

[11] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC*, 2003.

[12] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[13] F. Perronnin, G. Dance, C.and Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *ECCV*, 2006.

[14] 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.

[15] J. Sivic, B. Russell, A. Efros, A. Zisserman, and B. Freeman. Discovering objects and their location in images. In *ICCV*, 2005.

[16] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[17] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Describing visual scenes using transformed dirichlet processes. In *Advances in Neural Information Processing Systems 18*. 2006.

[18] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.