Discovering object categories in image collections *

Josef Sivic[1] Bryan C. Russell[2] Alexei A. Efros[3] Andrew Zisserman[1] William T. Freeman[2]

[1] Dept. of Engineering Science
University of Oxford
Oxford, OX1 3PJ, U.K.
{josef,az}@robots.ox.ac.uk

[2] CS and AI Laboratory
Massachusetts Institute of Technology
MA 02139, Cambridge, U.S.A.
{brussell,billf}@csail.mit.edu

[3] School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, U.S.A
efros@cs.cmu.edu

November 1, 2004, Updated: February 18, 2005

Abstract

Given a set of images containing multiple object categories, we seek to discover those categories and their image locations without supervision. We achieve this using generative models from the statistical text literature: probabilistic Latent Semantic Analysis (pLSA), and Latent Dirichlet Allocation (LDA). In text analysis these are used to discover topics in a corpus using the bag-of-words document representation. Here we discover topics as object categories, so that an image containing instances of several categories is modelled as a mixture of topics.

The models are applied to images by using a visual analogue of a word, formed by vector quantizing SIFT like region descriptors. We investigate a set of increasingly demanding scenarios, starting with image sets containing only two object categories through to sets containing multiple categories (including airplane, cars, faces, motorbikes, spotted cats) and background clutter. The object categories sample both intra-class and scale variation, and both the categories and their approximate spatial layout are found without supervision.

We also demonstrate classification of unseen images and images containing multiple objects. Performance of the proposed unsupervised method is compared to the semi-supervised approach of [7].

1. Introduction

Common approaches to object recognition involve some form of supervision. This may range from specifying the object's location and segmentation, as in face detection [17, 24], to providing only auxiliary data indicating the object's identity [1, 5, 7, 25]. For a large dataset, any annotation is expensive, or may introduce unforeseen biases. Results in speech recognition and machine translation highlight the importance of huge amounts of training data. The quantity of good, unsupervised training data – the set of still images – is orders of magnitude larger than the visual data available with annotation. Thus, one would like to observe many images and infer models for the classes of visual objects contained within them *without* supervision. This raises the scientific question which, to our knowledge, has not been convincingly answered before: Is it possible to learn visual object classes simply from looking at images?

Given large quantities of training data there has been notable success in unsupervised topic discovery in text, and it is this success that we wish to build on. We apply models used in statistical natural language processing to discover object categories and their image layout analogously to topic discovery in text. Documents are images and we quantize local appearance descriptions to form visual "words" [4, 18, 20, 26]. The two models we investigate are the probabilistic Latent Semantic Analysis (pLSA) of Hofmann [9, 10], and the Latent Dirichlet Allocation (LDA) of Blei *et al.* [3]. Both use the 'bag of words' model, where positional relationships between features are ignored. This greatly simplifies the analysis, since the data are represented by an observation matrix, a talley of the counts of each word (rows) in every document (columns).

The 'bag of words' model offers a rather impoverished representation of the data because it ignores any spatial relationships between the features. Nonetheless, it has been surprisingly successful in the text domain, because of the high discriminative power of some words and the redundancy of language in general. But can it work for images, where the spatial layout of the features is almost as important as the features themselves? While it seems implausible, there are several reasons for optimism: (i) as opposed to old corner detectors, modern feature descriptors have become powerful enough to encode very complex visual stimuli, making them quite discriminative; (ii) natural images are also very redundant (i.e. given a bag of features from an image, it is highly unlikely to find another natural image with the same features); (iii) because features are allowed to overlap in the image, some spatial information is implicitly preserved (i.e. randomly shuffling bits of the image around will almost cer-

*Published as MIT AI Memo: AIM-2005-005

tainly change the bag of words description). So, while these spatial relationships must eventually be taken into account, here we are investigating how far the bag of words model can be pushed in the image domain.

To use pLSA/LDA generative statistical models, we seek a vocabulary of visual words which will be insensitive to changes in viewpoint and illumination. We use vector quantized SIFT descriptors [12] computed on affine covariant regions [13, 14, 16]. Affine covariance gives us tolerance to viewpoint changes; SIFT descriptors, based on histograms of local orientation, give some tolerance to illumination change. Others have used similar descriptors for object classification [4, 15], but in a supervised setting.

We compare the two statistical models with a control global texture model, similar to those proposed for preattentive vision [22] and image retrieval [19]. Sect. 2 describes the pLSA and LDA statistical models; various implementation details are given in Sect. 3. To explain and compare performance, we apply the models to a series of progressively more challenging datasets of visual images in Sect. 4. We summarize in Sect. 5.

2. The pLSA and LDA models

We will describe the models here using the original terms 'documents' and 'words' as used in the text literature. Our visual application of these (as images and visual words) is then given in the following sections.

Suppose we have N documents containing words from a vocabulary of size M. The corpus of text documents is summarized in a M by N co-occurrence table N, where $n(w_i, d_j)$ stores the number of occurrences of a word w_i in document d_j. This is the bag of words model. In addition, there is a hidden (latent) topic variable z_k associated with each occurrence of a word w_i in a document d_j.

pLSA: The joint probability $P(w_i, d_j, z_k)$ is assumed to have the form of the graphical model shown in figure 1(a). Marginalizing over topics z_k determines the conditional probability $P(w_i|d_j)$:

$$P(w_i|d_j) = \sum_{k=1}^{K} P(z_k|d_j)P(w_i|z_k), \qquad (1)$$

where $P(z_k|d_j)$ is the probability of topic z_k occurring in document d_j; and $P(w_i|z_k)$ is the probability of word w_i occurring in a particular topic z_k.

The model (1) expresses each document as a convex combination of K topic vectors. This amounts to a matrix decomposition as shown in figure 1(b) with the constraint that both the vectors and mixture coefficients are normalized to make them probability distributions. Essentially, each document is modelled as a mixture of topics – the histogram for a particular document being composed from a mixture of the histograms corresponding to each topic.

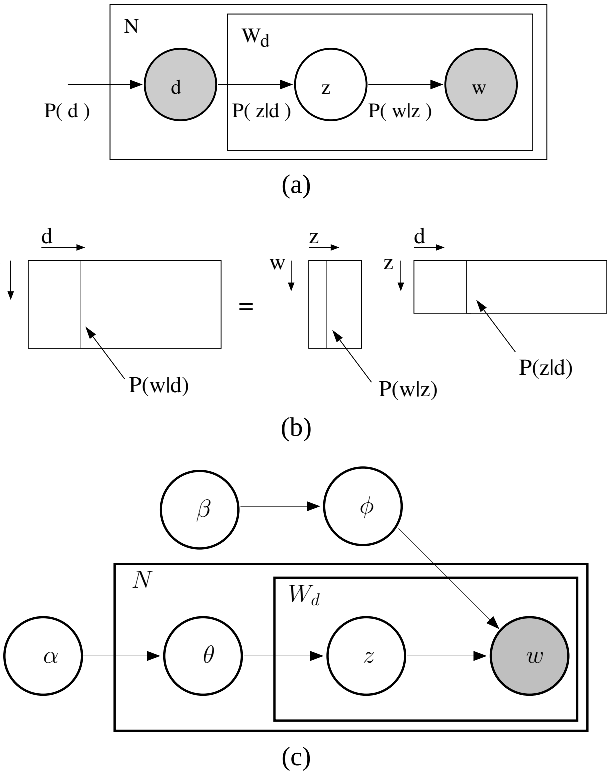


Figure 1: (a) pLSA graphical model, see text. Nodes inside a given box (plate notation) indicate that they are replicated the number of times indicated in the top left corner. Filled circles indicate observed random variables; unfilled are unobserved. (b) In pLSA the goal is to find the topic specific word distributions $P(w|z_k)$ and corresponding document specific mixing proportions $P(z|d_j)$ which make up the document specific word distribution $P(w|d_j)$. (c) LDA graphical model.

Fitting the model involves determining the topic vectors which are common to all documents and the mixture coefficients which are specific for each document. The goal is to determine the model that gives high probability to the words that appear in the corpus, and a maximum likelihood estimation of the parameters is obtained by maximizing the objective function:

$$L = \prod_{i=1}^{M}\prod_{j=1}^{N} P(w_i|d_j)^{n(w_i,d_j)}, \qquad (2)$$

where $P(w_i|d_j)$ is given by (1).

This is equivalent to minimizing the Kullback-Leibler divergence between the measured empirical distribution $\tilde{P}(w|d)$ and the fitted model. The model is fitted using the Expectation Maximization (EM) algorithm as described in [10].

LDA: In contrast to pLSA, LDA treats the multinomial weights over topics as latent random variables. The pLSA model is extended by sampling those weights from a Dirichlet distribution, the conjugate prior to the multinomial distribution. This extension allows the model to assign probabilities to data outside the training corpus and uses fewer parameters, thus reducing overfitting (see [3] for a detailed comparison). The LDA model is shown in Figure 1(c), where W_d is the number of words in document d. The goal is to maximize the following likelihood:

$$p(\boldsymbol{w}|\phi,\alpha,\beta) = \int \sum_{\boldsymbol{z}} p(\boldsymbol{w}|\boldsymbol{z},\phi)p(\boldsymbol{z}|\theta)p(\theta|\alpha)p(\phi|\beta)d\theta \tag{3}$$

where θ and ϕ are multinomial parameters over the topics and words respectively and $p(\theta|\alpha)$ and $p(\phi|\beta)$ are Dirichlet distributions parameterized by the hyperparameters α and β. Since the integral is intractable to solve directly, we solve for the parameters using Gibbs sampling, as described in [8].

The hyperparameters control the mixing of the multinomial weights (lower values give less mixing) and can prevent degeneracy. As in [8], we specialize to scalar hyperparameters (e.g. $\alpha_i = a \forall i$). For this paper, we used $\alpha_i = 0.5$ and $\beta_j = 0.5$.

3. Implementation details

Obtaining visual words: Two types of affine co-variant regions are computed for each image. The first is constructed by elliptical shape adaptation about an interest point. The method is described in [14, 16]. The second is constructed using the maximally stable procedure of Matas *et al.* [13] where areas are selected from an intensity watershed image segmentation. For both of these we use the binaries provided at [23]. Both types of regions are represented by ellipses. These are computed at twice the originally detected region size in order for the image appearance to be more discriminating.

Each ellipse is mapped to a circle by appropriate scaling along its principal axes and a SIFT descriptor computed. There is no rotation of the patch. Alternatively, the SIFT descriptor could be computed relative to the the dominant gradient orientation within a patch, making the descriptor rotation invariant [12]. The SIFT descriptors are then vector quantized into the visual 'words' for the vocabulary. The vector quantization is carried out here by k-means clustering computed from about 300K regions. The regions are those extracted from a random subset (about one third of each category) of images of airplanes, cars, faces, motorbikes and backgrounds (see experiment **(E)** in section 4). About 1K clusters are used for each of the Shape Adapted and Maximally Stable regions, and the resulting total vocabulary has 2,237 words. The number of clusters, k, is clearly an important parameter. The intention is to choose the size of k to determine words which give some intra-class generalization. Two examples of visual words are shown in Fig. 2.

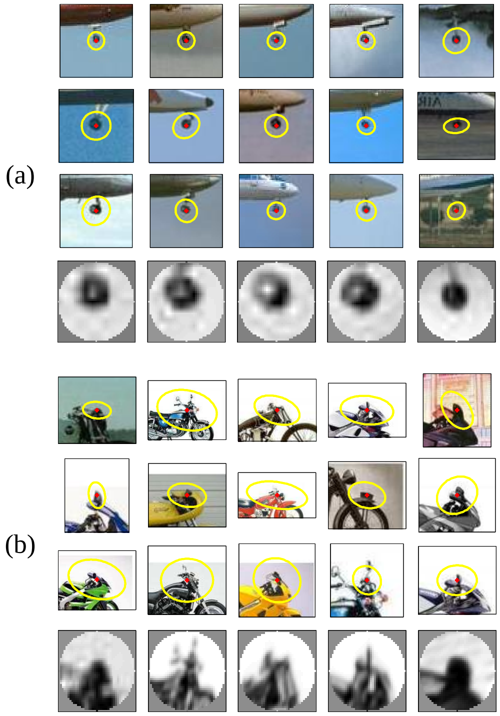


Figure 2: Two examples of visual words. (a) A wheel of an airplane. (b) A motorbike handle. In each case, the top three rows show 15 occurrences of this visual word in different images with the elliptical region superimposed. The bottom row shows affine normalized regions for the top row of images. Note that the normalized regions appear quite similar – which is why they are grouped in the same cluster. In the original images, the elliptical regions exhibit intra-class variation, and varying scale (the scaling is removed in this display as the ellipses are size normalized for visibility).

In text, a word with two different meanings is called polysemous (e.g. 'bank' as in (i) a money keeping institution, or (ii) a river side). We observe the visual analogue of polysemy in figure 3. However, the generative models are designed to cope with polysemous words. Such a word would have a high probability in two different topics. The hidden topic variable associated with each word occurrence in a particular document can assign such a word to particular topic depending on the context of the document. We return to this point in section 4.2.

Global Texture Model: To understand what level of performance can be accounted for by low-level image process-

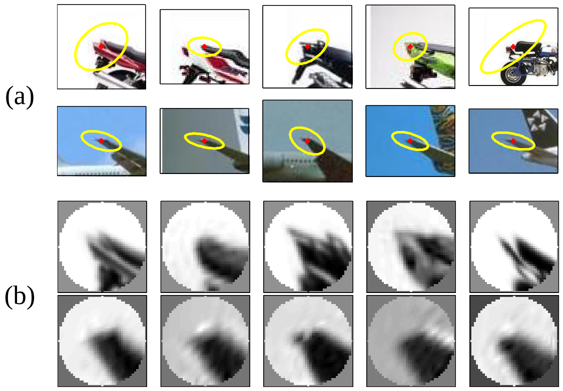


Figure 3: **Polysemy.** Example of a single visual word corresponding to two different (but locally similar) parts on two different object categories. (a) Top row shows occurrences of this visual word on the motorbike category, bottom row on the airplane category. The parts tend to occur consistently on different categories, i.e. this visual word fires mostly on the motorbike saddle and the airplane wing. (b) Corresponding normalized frames. Note the similarity of the normalized patches.

ing, we also implemented a simple texture clustering algorithm as our baseline method. The algorithm computes global feature histograms for each image in the database, and then clusters these histograms, again using k-means.

Using color histograms will not work for computing object classes since individual objects in each class will not necessarily have the same colors. Using texture histograms, on the other hand, does capture quite a lot about many types of objects (e.g. buildings have many vertical and horizontal edges). We experimented with several ways to represent texture and found the following simple method to yield the best results.

For each image in the database (grayscale), we compute the magnitude and the orientation of the gradient at each pixel. The gradient magnitudes are collected into a histogram (10 bins), and the corresponding orientations which are greater than a threshold (0.02 in our case) are also collected into a histogram (12 bins). The histograms are normalized and concatenated into a single 22-dimensional vector; one per each image in the database. The vectors are then clustered using k-means.

Model learning: In the case of pLSA, the EM algorithm is initialized randomly, typically converges in 100-300 iterations, and takes about 10 mins to run on 3K images (Matlab implementation on a 2GHz PC). For LDA, we use Gibbs sampling to draw samples from the posterior $p(z_i|\boldsymbol{z}_{-i}, \boldsymbol{w}, \alpha, \beta)$ over topics, where $\boldsymbol{z}_{-i}$ indicates all other topic variables except z_i. We do this for 50 rounds and use the topic settings that maximizes the log-likelihood. Usually, we reach the maximum within the first 30 rounds. The topic settings can then be used to compute relevant parameters, such as ϕ and θ. This process takes on the order of an hour for 3K images. However, the independences in the model allows for parallelism to be exploited.

4. Topic Discovery Experiments

Given a collection of completely unlabelled images, our goal is to automatically discover the visual categories present in the data and localize them in the image. To this end, we carry out a set of quantitative experiments with progressively increasing level of visual difficulty. Since here we know the object instances in each image, we use this information as a performance measure. A confusion matrix is computed for each experiment for each of the three models being tested (pLSA, LDA, and the baseline texture model). Below we describe the datasets (1-8), the experiments (A-F), and summarize the results for each.

Data sets: Our data set consists of six categories from the Caltech image datasets (as previously used by Fergus *et al.* [7] for semi-supervised classification), and two categories ((7) and (8) below) from the more difficult 101 category dataset [6].

Label	description	# images
(1)	All faces	435
(1ub)	Faces on uniform background a cropped version of (1)	435
(2)	All motorbikes	800
(2ub)	Motorbikes on uniform background a subset of (2)	349
(3)	All airplanes	800
(3ub)	Airplanes on uniform background a subset of (3)	263
(4)	Cars rear	1155
(5)	Leopards	200
(6)	Background	1370
(7)	Watch	241
(8)	Ketch	114

The reason for picking these particular categories is pragmatic: they are the ones with the greatest number of images per category. All images have been converted to grayscale before processing. Otherwise they have not been altered in any way, with one notable exception: a large number of images in the motorbike category (2) and airplane category (3) have a white border around the image which we have removed since it was providing an artifactual cue for object class.

4.1. Classification

In the following we carry out a series of experiments varying the number and difficulty of the categories. In each case images are pooled from a number of original datasets, and the three models are fitted to the ensemble of images (with

Ex	Categories	pLSA		LDA		Texture	
		%	#	%	#	%	#
A	2,3 ub	100	1	99	7	91	53
B	1-3 ub	100	2	96	40	94	55
C	1-3	97	56	96	71	91	170
D	1-4	98	70	87	365	72	1060
E	1-4 + bg	78	931	77	970	73	1174
E6*	1-4 + bg	76	1072	–	–	–	–
E7*	1-4 + bg	83	768	–	–	–	–
F	1-5,7-8 + bg	59	1515	64	1458	47	2093

Figure 4: Summary of the experiments. Column '%' shows the classification accuracy measured by the average of the diagonal of the confusion matrix. Column '#' shows the total number of misclassifications. See text for a more detailed description of the experimental results. (*) In the case of E6/E7 the two/three background topics are classified as one category.

no knowledge of the image's labels) for a specified number of topics, K. This is by default set equal to the number of categories in the dataset. For example, in experiment (B) the images are pooled from three categories (airplanes, cars and motorbikes, all with uniform backgrounds) and models with $K = 3$ objects (topics) are fitted. In the case of pLSA, the model determines the mixture coefficients $P(z_k|d_j)$ for each image (document) d_j (where $z \in \{z_1, z_2, z_3\}$ for the three topics in example (C)). An image d_j is then classified as containing object k according to the maximum of $P(z_k|d_j)$ over k. In the LDA case, we classify based on the topic mixture weights θ, which can be computed using the samples drawn by the Gibbs sampler.

We performed the following experiments. The results are summarized in table 4.

(A) Two object categories (2ub,3ub), uniform backgrounds. This is a relatively easy test (Airplanes vs. Motorbikes) with no background clutter to worry about. Both models perform very well with pLSA having only 1 misclassified image, and LDA only 7.

The baseline texture model has 53 misclassified. Although not perfect, the simple texture model performs surprisingly well. This is probably due to the fact that there are only two categories which can be easily separated by gradient orientations (airplanes are less textured and more recto-linear, while motorbikes have more texture at all orientations).

(B) Three object categories (1ub,2ub,3ub), uniform backgrounds. Here we increase the number of categories by one, adding the cropped face dataset. The performance of all three models is similar to that of experiment (A), despite the addition of over 250 more images.

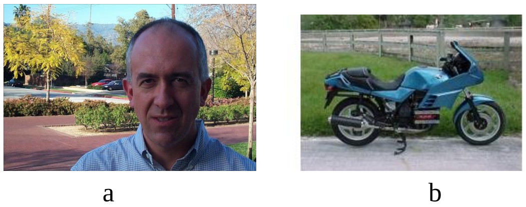


Figure 5: (a) Example of a face image which is classified in experiment C as a motorbike due to similar background (grass, trees) to many motorbikes images. (b) An example of a motorbike with mostly grass background.

(C) Three object categories (1,2,3), cluttered backgrounds . Now we introduce a cluttered background with each object class. Because the backgrounds are somewhat correlated with the objects, we expect similar results regarding the numbers of topics discovered, with the correlated backgrounds being considered parts of the objects. However, this is a more visually challenging task as the background is not a single object, but much more varied and disparate. The results for pLSA are summarized in the following confusion table. LDA exhibits a very similar behavior.

True Class →	Faces	Motorb	Airplan
Topic 1 - Faces	95.17	0.25	0.75
Topic 2 - Motorb	4.83	99.12	2.75
Topic 3 - Airplan	0.00	0.62	96.50

It is interesting to examine the images that are confused between the topics. Essentially the confusion arises because the background is in common between the images, see figure 5. This motivates experiment (E) where background images are added, and there is the opportunity for the models to discover the background as an object.

(D) Four object categories (1,2,3,4), cluttered backgrounds. Here, we add a fourth category (Cars rear), all with cluttered backgrounds and significant scale variations. An interesting observation comes from varying the number of topics, K. In the case of $K = 4$, we discover the four different categories in the dataset with very high accuracy (see table 4). In the case of $K = 5$, the car dataset splits into two subtopics. This is because the data contains sets of many repeated images of the same car. Increasing K to 6 splits the motorbike data into sets with a plain background and cluttered background similar to our manual split of the data for experiments A and B. Increasing K further to 7 and 8 'discovers' two more sub-groups of car data containing again other repeated images of the same/similar cars.

It is also interesting to see the visual words which are most probable for an object, by selecting those with high topic specific probability $P(w_i|z_k)$. These are shown for the pLSA model for the case of $K = 4$ in figure 6.

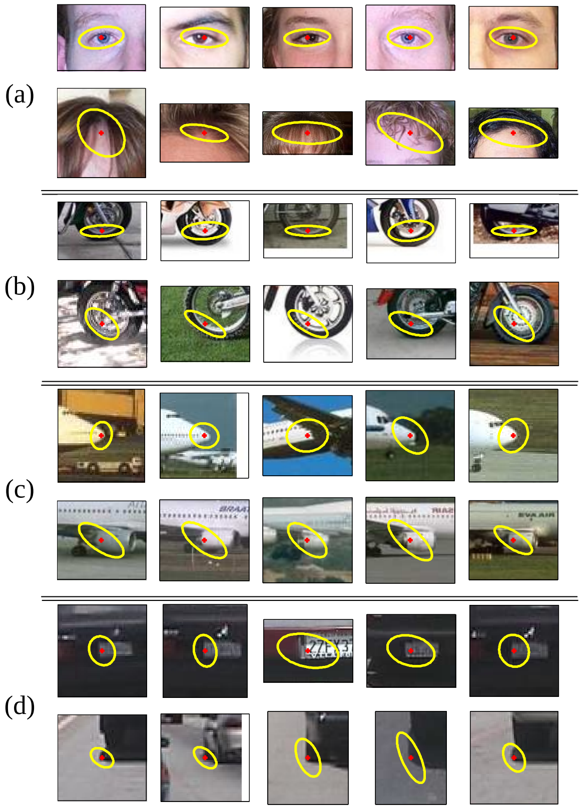


Figure 6: Two most likely words (shown by 5 examples in a row) for four learnt topics in experiment (D): (a) Faces, (b) Motorbikes, (c) Airplanes, (d) Cars.

(E) Four object categories (1,2,3,4) plus "background" category (6). Here we add an explicit "background" category (indoor and outdoor scenes around Caltech campus) to our experiment D. The reason for adding these additional images is to give the methods the opportunity of discovering background "objects".

The confusion tables for the three methods are shown as images in figure 7(a). It is evident, for example, that for both pLSA and LDA the first topic confuses faces and backgrounds to some extent.

We now carry out further pLSA model fits with $K = 6, 7$. The result is very interesting: the confusion between the four object categories decreases significantly, and instead the background is treated as three separate topics, see figure 7(b). Because the background is so varied, it is being treated as three distinct objects, roughly corresponding to local feature-like texture, building/office-like texture and stochastic-like texture. Examples of visual words with high probability under these background topics/objects are given in figure 8. This example reiterates that an image is being described as a mixture over topics. Examining the posteriors, it can be seen that a typical image consisting of foreground object (e.g. a motorbike) and background is now described as a mixture of motorbike and the background topics (e.g. texture). We return to this point below in section 4.2.

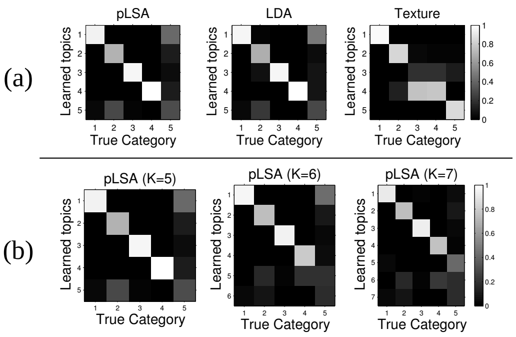


Figure 7: (a) Confusion tables for pLSA, LDA and texture for experiment (E) with 5 learned topics. Brightness indicates number. The ideal is bright down the diagonal. (b) Confusion tables for pLSA for increasing number of topics (K=5,6,7) respectively. Note how the background (category 5 splits into 2 and 3 topics (for K=6 and 7 respectively) and that some amount of the confusion between categories and background is removed.

(F) Seven object categories (1,2,3,4,5,7,8) plus "background" category (6). In our biggest experiment, we used all our datasets with real backgrounds (adding Leopards, Watch and Ketch to previous experiments). Even though the new categories all had substantially fewer images (around 200), the results are still encouraging.

Discussion: In the experiments it was necessary to specify the number of topics K, however Bayesian [21] or minimum complexity methods [2] can be used to infer the number of topics implied by a corpus.

While designing these experiments, we grew to appreciate the many difficulties in searching for good datasets. Finding a collection of images containing objects of the same category that is large enough (at least 200 images), hard enough (good intra-class variation between objects), but doable (the intra-class variation is based on appearance, not semantics) is not an easy task! Dealing with realistic backgrounds presents another set of issues. In a full object-recognition system there should not be anything called "background" – every object in the scene must be explained. However, this requires enough training data to cover each object *independently of all the others*. For example, if all the airplanes in the dataset are pictured on tarmacs (with no airplanes in the air, and no empty tarmacs), then there is no way for the system to learn that these are actually two distinct object classes. In these experiments, we addressed these problems in two ways: (i) every attempt was made to use datasets with varied backgrounds, (ii) a

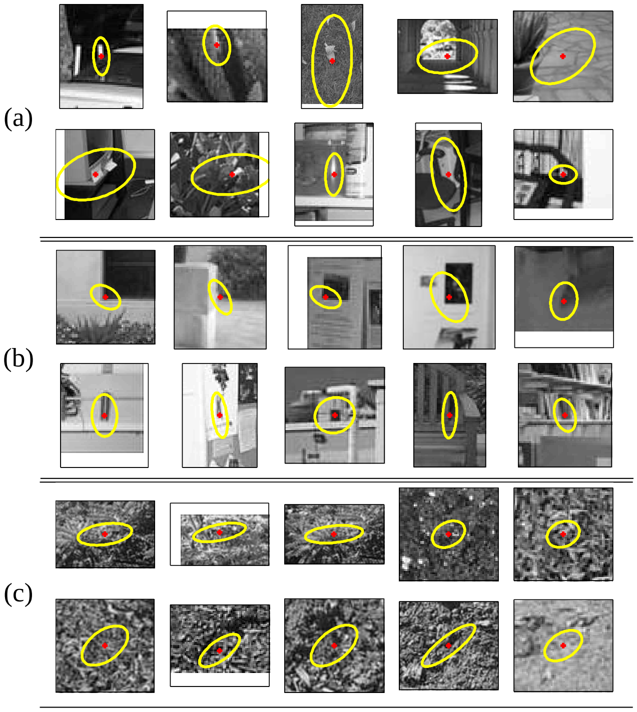


Figure 8: Two most likely words (shown by 5 examples in a row) for the three background topics learned in experiment E: (a) topic 2,mainly local feature-like structure (b) topic 4, mainly corners and edges coming from the office/building scenes, (c) topic 5, mainly textured regions like grass and trees. For topic numbers refer to figure 11(c).

"background" category was added, with backgrounds similar to the ones used in other categories.

Classifying unseen images: The learned topics can also be used for classifying unseen images, a task similar to the one in Fergus *et al.* [7]. In the case of pLSA, the topic specific distributions $P(w|z)$ are learned from a separate set of 'training' images. When observing a new *unseen* 'test' image, the document specific mixing coefficients $P(z|d_{test})$ are computed using the 'fold-in' heuristic described in [9]. In particular, the unseen image is 'projected' on the simplex spanned by learned $P(w|z)$, i.e. the mixing coefficients $P(z_k|d_{test})$ are sought such that the Kullback-Leibler divergence between the measured empirical distribution $\tilde{P}(w|d_{test})$ and $P(w|d_{test}) = \sum_{k=1}^{K} P(z_k|d_{test})P(w|z_k)$ is minimized. This is achieved by running EM in a similar manner to that used in learning, but now only the coefficients $P(z_k|d_{test})$ are updated in each M-step. The learned $P(w|z)$ are kept fixed.

To compare performance with Fergus *et al.* [7], experiment E was modified such that only the 'training' subsets for each category (and all background images) from [7] were used to fit the pLSA model with 7 topics (four object topics and three background topics). The 'test' images from [7] were than 'folded in' as described above. In the first test the confusion between different object categories is examined. Each test image is assigned to object topic k with maximum $P(z_k|d_{test})$ (background topics are ignored here). The confusion table is shown in figure 9.

True Class →	Faces	Motorb	Airplan	Cars rear
Topic 1 - Faces	99.54	0.25	1.75	0.75
Topic 2 - Motorb	0.00	96.50	0.25	0.00
Topic 3 - Airplan	0.00	1.50	97.50	0.00
Topic 4 - Cars rear	0.46	1.75	0.50	99.25

Figure 9: Confusion table for unseen test images (modified experiment E). Note there is very little confusion between different categories. See text.

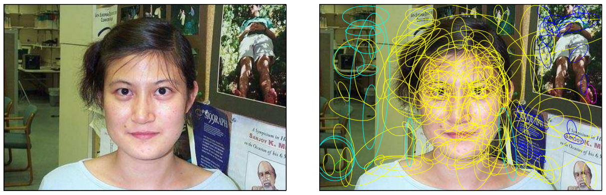
Object categ.	pLSA (a)	pLSA (b)	Fergus *et al.* [7]
Faces	5.3	3.3	3.6
Motorbikes	15.4	8.0	6.7
Airplanes	3.4	1.6	7.0
Cars rear*	21.4 / 11.9	16.7 / 7.0	9.7

Figure 10: Equal error rates for image classification task for pLSA and the method of [7]. Test images of a particular category were classified against (a) testing background images (test performed in [7]) and (b) testing background images *and* testing images of all other categories. The improved performance in (b) is because our method exhibits very little confusion between different categories. (*) The two performance figures correspond to training on 400 / 900 background images respectively. In both cases, classification is performed against an unseen test set of road backgrounds (as in [7]), which was folded-in. See text for explanation.

In the second test we examine performance in classifying (unseen) images against (unseen) background images. The pLSA model is fitted to training subsets of each category and a training subset of only 400 (out of 900) background images. Testing images of each category and testing background images are 'folded-in'. The mixing proportion $P(z_k|d_{test})$ for topic k across the testing images d_{test} (i.e. a row in the landscape matrix $P(z|d)$ in figure 1b) is then used to produce a ROC curve for the topic k. Equal error rates for the four object topics are reported in figure 10.

Note that for Airplanes and Faces our performance is similar to that of [7] despite the fact that our 'training' is unsupervised in the sense that the identity of the object in an image is *not known*. This is in contrast to [7], where each image is labelled with an identity of the object it contains, i.e. about 5×400 items of supervisory data vs. one label (the number of topics) in our case.

In the case of motorbikes we perform worse than [7] mainly due to confusion between motorbike images containing textured background and textured background topic (similar problem is shown in figure 5). The performance on

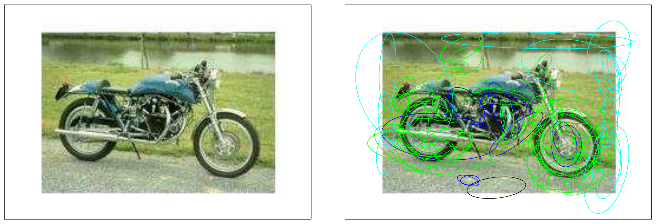


(a) (b)

Topic	$P(topic \vert image)$	# regions
1 Motorbikes (green)	0.07	1
2 Backg I (magenta)	0.09	1
3 Face (yellow)	0.48	128
4 Backg II (cyan)	0.17	12
5 Backg III (blue)	0.15	23
6 Cars (red)	0.03	0
7 Airplane (black)	0.00	0

(c)

Figure 11: Image as a mixture of visual topics (Experiment E) - I. (a) Original frame. (b) Image as a mixture of a face topic (yellow) and background topics (blue, cyan). Only elliptical regions with topic posterior $P(z|w,d)$ greater than 0.8 are shown. In total 7 topics were learned for this dataset which contained (faces, motorbikes, airplanes, cars, and background images). The other topics are not significantly present in the image since they mostly represent the other categories and other types of background. Table (c) shows the mixture coefficients $P(z|d)$ for this particular image. In total there are 693 elliptical regions in this image of which 165 (102 unique visual words) have $P(z|w,d)$ above 0.8 (those shown in (b)).



(a) (b)

Figure 12: Image as a mixture of visual topics II. For this image $P(z|d)$ is 0.48 (motorbikes - green), 0.01 (bg I - magenta), 0.04 (face - yellow), 0.20 (bg II - cyan), 0.14 (bg III - blue), 0.02 (cars - red), 0.10 (airplane - black). In total there are 466 elliptical regions in this image of which 102 (80 unique visual words) have $P(z|d)$ above 0.8 (those shown in (b)).

Cars rear is poor because Car images are split between two topics in training (a similar effect happens in experiment D for K=6). This splitting can be avoided by including more background images. In order to make results comparable with [7], Cars rear images were classified against completely new background dataset containing mainly empty roads. This dataset was not seen in the learning stage and had to be 'folded-in' which makes the comparison on Cars rear slightly unfair.

4.2. Segmentation

In this section we evaluate the image's spatial segmentation that have been discovered by the model fitting. As a first thought, it is absurd that a bag of words model could possibly have anything useful to say about image segmentation, since all spatial information has been thrown away. However, the pLSA model delivers the posteriors

$$P(z_k|w_i,d_j) = \frac{P(w_i|z_k)P(z_k|d_j)}{\sum_{l=1}^{K} P(w_i|z_l)P(z_l|d_j)}, \qquad (4)$$

and consequently for a word occurrence in a particular document we can examine the probability of different topics.

Figures 11 and 12 show examples of 'topic segmentation' induced by $P(z_k|w_i,d_j)$ for the case of experiment (E) with 7 topics. Figure 13 shows examples of topic segmentation for unseen images (modified experiment E7). In particular, we show only visual words with $P(z_k|w_i,d_j)$ greater than 0.8. There is an impressive alignment of the words with the corresponding object areas of the image. Note the words shown are not simply those most likely for that topic. Rather, from (4), they have high probability of that topic *in this image*. This is an example of overcoming polysemy – the probability of the particular word depends not only on the probability that it occurs within that topic (face, say) but also on the probability that the face topic has for that image, i.e. the evidence for the face topic from other regions in the image.

There are no examples of images containing several different objects in the datasets we are using, so we provide a small number of additional images. In the pLSA case we have added these images to the dataset of experiment E and re-fitted the model to all images. In the case of LDA, we use the learned parameters to do prediction on the unseen mixed-category images. An example of segmentations for an image containing a car and a motorbike is shown in figure 14(a) and a car and a face in 14(b). The segmentations were obtained in the same way as described above. Note in the case of pLSA it was necessary to refit the model to all the images (although this can be avoided [10]), however in the case of LDA the mixture weights could be inferred directly.

5. Conclusions

We have demonstrated that it is possible to learn visual object classes simply by looking; in experiments (A) through (F) we identify the object categories for each image with the high reliabilities shown in figure 4, using a corpus of unlabelled images. Furthermore, the visual words with the highest posterior probabilities for each object correspond fairly well to the spatial locations of each object. This is rather remarkable considering our use of the bag of words

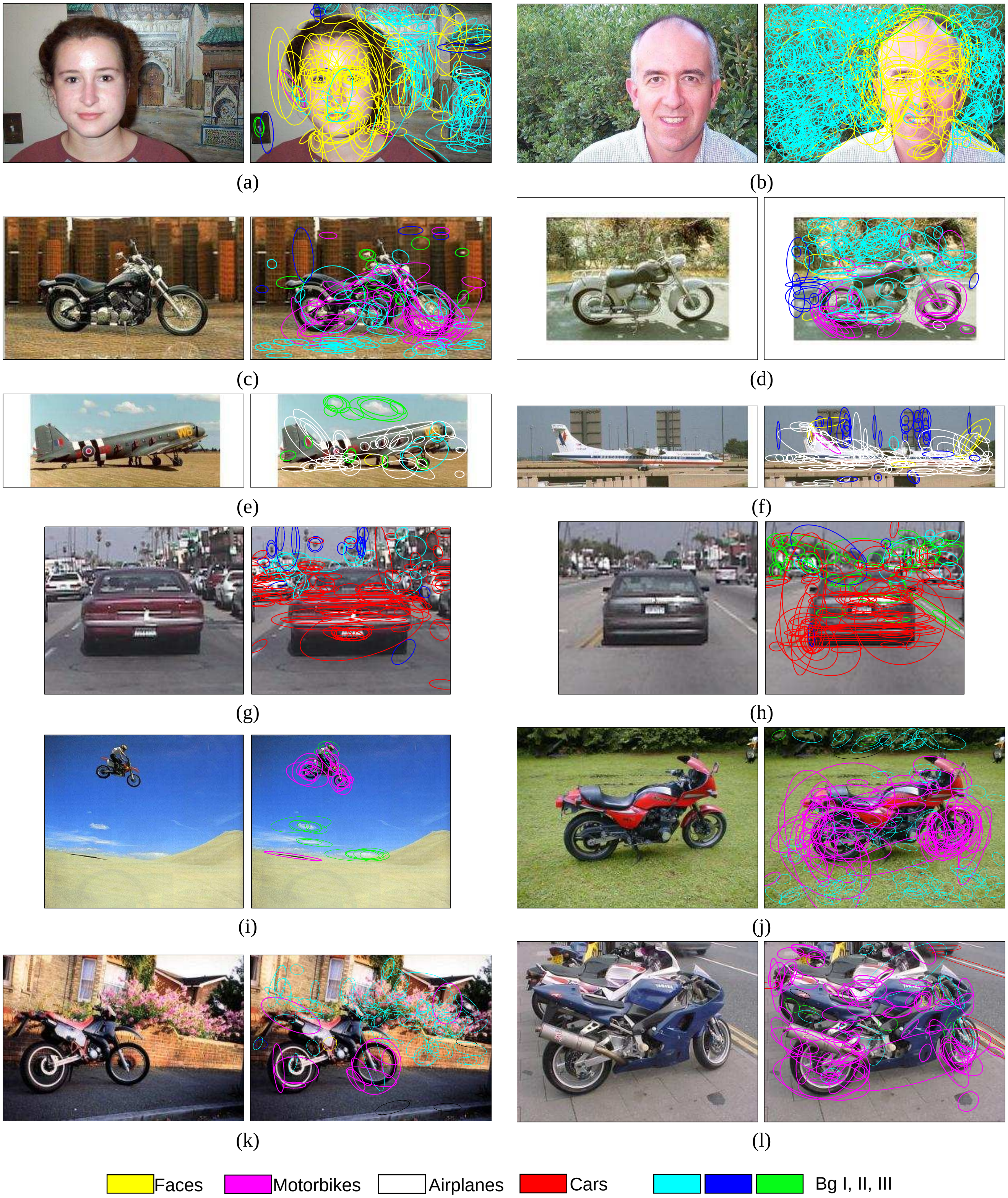
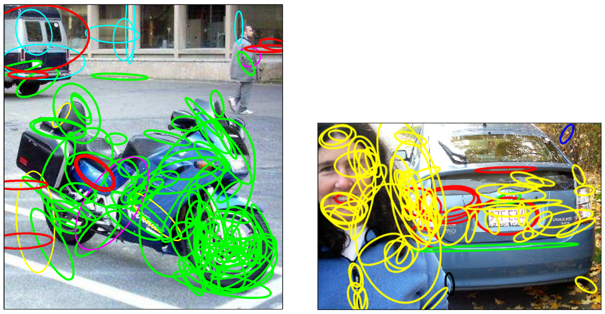


Figure 13: Examples of segmentation for new unseen images using pLSA. Topics were learned from a set of training images from each category and all background images. Each new image was 'folded in', see text. Two examples from each category are shown: Faces (a,b), Motorbikes (c,d), Airplanes (e,f), Cars rear (g,h). Four examples from the ETH motorbike dataset [11] are shown in (i–l). Note the significant changes in scale and viewpoint. The color key for the seven different topics is shown at the bottom of the figure.



(a) (b)

Figure 14: Multiple objects in an image. **(a) pLSA example:** Two objects are present in this image: a motorbike (topic 1 - green) and a car (topic 6 - red). The learned mixture coefficients $P(z|d)$ are 0.41 (motorbikes - green), 0.02 (bg I - magenta), 0.16 (face - yellow), 0.19 (bg II - cyan), 0.04 (bg III - blue), 0.14 (cars - red), 0.02 (airplane - black). In total there are 740 elliptical regions in this image of which 95 (72 unique visual words) are shown (have $P(z|w, d)$ above 0.8). **(b) LDA example:** Two objects are present in this image. a face (yellow) and a car (red). The learned mixing weights θ are 0.19 car (red), 0.07 motorbike (green), 0.16 airplane (black), 0.14 background (blue), 0.44 face (yellow).

model.

We have explored an extreme approach, no spatial propagation of information, and have met with surprising success. The current work provides a foundation for spatial inference: the posterior marginal probabilities for object membership for each local region and each object. Already, we have shown these probabilities to be useful for identifying and localizing objects that have been discovered from a training corpus. We expect they will also prove useful for tasks such as combining topic discovery with spatial inference and perceptual organization, or image retrieval. For example, the topic vectors we have discovered may now be applied directly as 'semantic vectors' for retrieval from image databases, and we anticipate significant performance improvements compared to standard approaches such as LSA [20].

References

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

[2] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Trans. on Information Therory*, 4:1034–1054, 1991.

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[5] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. ECCV*, 2002.

[6] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE CVPR Workshop of Generative Model Based Vision*, 2004.

[7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. CVPR*, 2003.

[8] T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

[9] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1999.

[10] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 43:177–196, 2001.

[11] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

[12] D. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, pages 1150–1157, 1999.

[13] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. BMVC.*, pages 384–393, 2002.

[14] K. Mikolajczyk and C. Schmid. An affine invariant interest point detector. In *Proc. ECCV*. Springer-Verlag, 2002.

[15] A. Opelt, A. Fussenegger, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proc. ECCV*, 2004.

[16] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *Proc. ECCV*, volume 1, pages 414–431. Springer-Verlag, 2002.

[17] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *Proc. CVPR*, 2000.

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

[19] J. Smith and S. F. Chang. Automated image retrieval using color and texture. *IEEE PAMI*, 1996.

[20] F. Souvannavong, B. Merialdo, and B. Huet. Improved video content indexing by multiple latent semantic analysis. In *Proc. CIVR*, 2004.

[21] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. In *Proc. NIPS*, 2004.

[22] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):153–167, 2003.

[23] `http://www.robots.ox.ac.uk/˜vgg/research/affine/`.

[24] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, 2001.

[25] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. ECCV*, pages 18–32, 2000.

[26] R. Zhang and Z. Zhang. Hidden semantic concept discovery in region based image retrieval. In *Proc. CVPR*, 2004.