Discovering objects and their location in images

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

Abstract

We seek to discover the object categories depicted in a set of unlabelled images. We achieve this using a model developed in the statistical text literature: probabilistic Latent Semantic Analysis (pLSA). In text analysis this is used to discover topics in a corpus using the bag-of-words document representation. Here we treat object categories as topics, so that an image containing instances of several categories is modeled as a mixture of topics.

The model is applied to images by using a visual analogue of a word, formed by vector quantizing SIFT-like region descriptors. The topic discovery approach successfully translates to the visual domain: for a small set of objects, we show that both the object categories and their approximate spatial layout are found without supervision. Performance of this unsupervised method is compared to the supervised approach of Fergus et al. [8] on a set of unseen images containing only one object per image.

We also extend the bag-of-words vocabulary to include 'doublets' which encode spatially local co-occurring regions. It is demonstrated that this extended vocabulary gives a cleaner image segmentation. Finally, the classification and segmentation methods are applied to a set of images containing multiple objects per image. These results demonstrate that we can successfully build object class models from an unsupervised analysis of images.

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, 23], to providing only auxiliary data indicating the object's identity [1, 2, 8, 24]. 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 motivates 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. In our setting, documents are images and we quantize local appearance descriptors to form visual "words" [5, 19]. The two models we have investigated are the probabilistic Latent Semantic Analysis (pLSA) of Hofmann [9, 10], and the Latent Dirichlet Allocation (LDA) of Blei *et al.* [4, 7, 18]. Each model consistently gave similar results and we focus our exposition in this paper on the simpler pLSA method. Both models use the 'bag of words' representation, where positional relationships between features are ignored. This greatly simplifies the analysis, since the data are represented by an observation matrix, a tally 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 objects, where the spatial layout of the features may be 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) because visual features overlap in the image, some spatial information is implicitly preserved (i.e. randomly shuffling bits of the image around will almost certainly change the bag of words description). In this paper, we show that this optimism is not groundless.

While we ignore spatial position in our 'bag of words'

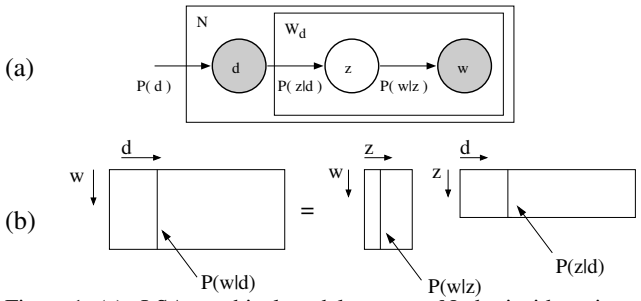

Figure 1: (a) pLSA graphical model, see text. Nodes inside a given box (plate notation) indicate that they are replicated the number of times indicated in the top left corner. Filled circles indicate observed random variables; unfilled are unobserved. (b) In pLSA the goal is to find the topic specific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

object class models, our models are sufficiently discriminative to localize objects within each image, providing an approximate segmentation of each object topic from the others within an image. Thus, these bag-of-features models are a step towards top-down segmentation and spatial grouping.

We take this point on segmentation further by developing a second vocabulary which is sensitive to the spatial layout of the words. This vocabulary is formed from spatially neighboring word pairs, which we dub *doublets*. We demonstrate that doublets provide a cleaner segmentation of the various objects in each image. This means that both the object category and image segmentation are determined in an unsupervised fashion.

Sect. 2 describes the pLSA statistical model; various implementation details are given in Sect. 3. To explain and compare performance, in Sect. 4 we apply the models to sets of images for which the ground truth labeling is known. We also compare performance with a baseline algorithm: a k-means clustering of word frequency vectors. Results are presented for object detection and segmentation. We summarize in Sect. 5.

2. The topic discovery model

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 $\mathtt{N}$, 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.

Fitting the model involves determining the topic vectors which are common to all documents and the mixture coefficients which are specific to 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].

3. Implementation details

Obtaining visual words: We seek a vocabulary of visual words which will be insensitive to changes in viewpoint and illumination. To achieve this we use vector quantized SIFT descriptors [11] computed on affine covariant regions [12, 13, 16]. Affine covariance gives tolerance to viewpoint changes; SIFT descriptors, based on histograms of local orientation, gives some tolerance to illumination change. Vector quantizing these descriptors gives tolerance to morphology within an object category. Others have used similar descriptors for object classification [5, 15], but in a supervised setting.

Two types of affine covariant regions are computed for each image. The first is constructed by elliptical shape adaptation about an interest point. The method is described in [13, 16]. The second is constructed using the maximally stable procedure of Matas *et al.* [12] where areas are selected from an intensity watershed image segmentation. For

both of these we use the binaries provided at [22]. 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 is computed. There is no rotation of the patch, i.e. the descriptors are rotation variant (alternatively, the SIFT descriptor could be computed relative to the the dominant gradient orientation within a patch, making the descriptor rotation invariant [11]). 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 Expt. (2) 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. This vocabulary is used for all the experiments throughout this paper.

In text analysis, 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 analogue of polysemy in our visual words, however, the topic discovery models can cope with these. A polysemous 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 a particular topic depending on the context of the document. We return to this point in section 4.3.

Doublet visual words: For the task of segmentation, we seek to increase the spatial specificity of object description while at the same time allowing for configurational changes. We thus augment our vocabulary of words with "doublets" – pairs of visual words which co-occur within a local spatial neighborhood. As candidate doublet words, we consider only the 100 words (or less) with high probability in each topic after an initial run of pLSA. To avoid trivial doublets (those with both words in the same location), we discard those pairs of ellipses with significant overlap. We then form doublets from all pairs of the remaining words that are within five nearest neighbors of each other. However, there is a preference for ellipses of similar sizes, and this is achieved by using a distance function (for computing the neighbors) that multiplies the actual distance (in pixels) between candidate ellipses by the ratio of the larger to smaller major axis of the two ellipses. Figure 2 illustrates the geometry and formation of the doublets. Figure 7d,e shows examples of doublets on a real image.

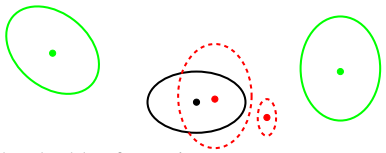

Figure 2: The doublet formation process. We search for the k nearest neighbors of the word ellipse marked in black, giving one doublet for each valid neighbor. We illustrate the process here with $k = 2$. The large red ellipse significantly overlaps the the black one and is discarded. After size-dependent distance scaling, the small red ellipse is further from the black center than the two green ellipses and is also discarded. The black word is thus paired with each of the two green ellipse words to form two doublets.

Model learning: For pLSA, the EM algorithm is initialized randomly and typically converges in 40–100 iterations. One iteration takes about 2.3 seconds on 4K images with 7 fitted topics and $\sim$300 non-zero word counts per image (Matlab implementation on a 2GHz PC).

Baseline method – k-means (KM): To understand the contributions of the topic discovery model to the system performance, we also implemented an algorithm using the same features of word frequency vectors for each image, but without the final statistical machinery. The standard k-means procedure is used to determine k clusters from the word frequency vectors, i.e. the pLSA clustering on KL divergence is replaced by Euclidean distance and each document is hard-assigned to exactly one cluster.

4. Experiments

Given a collection of unlabelled images, our goal is to automatically discover/classify the visual categories present in the data and localize them in the image. To understand how the algorithms perform, we train on image collections for which we know the desired visual topics.

We investigate three areas: (i) topic discovery – where categories are discovered by pLSA clustering on all available images, (ii) classification of unseen images – where topics corresponding to object categories are learnt on one set of images, and then used to determine the object categories present in another set, and (iii) object detection – where we also wish to determine the location and approximate segmentation of object(s) in each image.

We use two datasets of objects, one from Caltech [6, 8] and the other from MIT [21]. The Caltech datasets depict one object per image. The MIT dataset depicts multiple object classes per image. We report results for the three areas first on the Caltech images, and then in section 4.4 show their application to the MIT images.

Caltech image data sets. Our data set consists of images of five categories from the Caltech 101 datasets (as previously used by Fergus *et al.* [8] for supervised classification). The categories and their number of images are:

Ex	Categories	K	pLSA		KM baseline	
			%	#	%	#
(1)	4	4	98	70	72	908
(2)	4 + bg	5	78	931	56	1820
(2)*	4 + bg	6	76	1072	–	–
(2)*	4 + bg	7	83	768	–	–
(2)*	4 + bg-fxd	7	93	238	–	–

Table 1: 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 (2)* the two/three background topics are allocated to one category. Evidently the baseline method performs poorly, showing the power of the pLSA clustering.

faces, 435; motorbikes, 800; airplanes, 800; cars rear, 1155; background, 900. 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 and airplane category have a white border around the image which we have removed since it was providing an artifactual cue for object class.

4.1. Topic discovery

We carry out a sequence of increasingly demanding experiments. In each experiment images are pooled from a number of original datasets, and the pLSA and baseline models are fitted to the ensemble of images (with no knowledge of the image's labels) for a specified number of topics, K. For example, in Expt. (1) the images are pooled from four categories (airplanes, cars, faces and motorbikes) and models with $K = 4$ 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 $z \in \{z_1, z_2, z_3, z_4\}$ for the four topics). An image d_j is then classified as containing object k according to the maximum of $P(z_k|d_j)$ over k. This is essentially a one against many (the other categories) test. Since here we know the object instances in each image, we use this information as a performance measure. A confusion matrix is then computed for each experiment.

Expt. (1) Images of four object categories with cluttered backgrounds. The four Caltech categories have cluttered backgrounds and significant scale variations (in the case of cars rear). We investigate clustering as the number of topics is varied. In the case of $K = 4$, we discover the four different categories in the dataset with very high accuracy (see table 1). 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

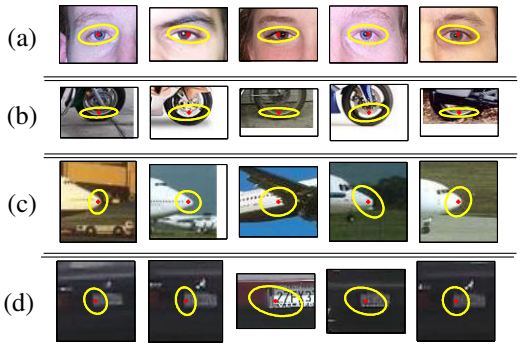

Figure 3: The most likely words (shown by 5 examples in a row) for four learnt topics in Expt. (1): (a) Faces, (b) Motorbikes, (c) Airplanes, (d) Cars.

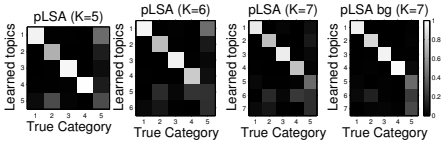

Figure 4: Confusion tables for Expt. (2) for increasing number of topics (K=5,6,7) and with 7 topics and fixed background respectively. Brightness indicates number. The ideal is bright down the diagonal. Note how the background (category 5) splits into 2 and 3 topics (for K=6 and 7 respectively) and that some amount of the confusion between categories and background is removed.

and cluttered background. 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 case of $K = 4$ in figure 3.

Thus, for these four object categories, topic discovery analysis cleanly separates the images into object classes, with reasonable behavior as the number of topics increases beyond the number of objects. The most likely words for a topic appear to be semantically meaningful regions.

Expt. (2) Images of four object categories plus "background" category. Here we add images of an explicit "background" category (indoor and outdoor scenes around Caltech campus) to the above Expt. (1). The reason for adding these additional images is to give the methods the opportunity of discovering background "objects".

The confusion tables as K is varied are shown as images in figure 4. It is evident, for example, that the first topic confuses faces and backgrounds to some extent. The results are summarized in table 1. The case of $K = 7$ with three topics allocated to the background gives the best performance. Examples of the most likely visual words for the three discovered background topics are shown in figure 5.

In the case of many of the Caltech images there is a strong correlation of the foreground and backgrounds (e.g.

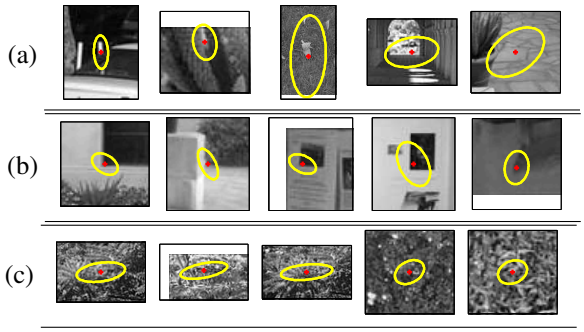

Figure 5: The most likely words (shown by 5 examples in a row) for the three background topics learned in Expt. (2): (a) Background I, mainly local feature-like structure (b) Background II, mainly corners and edges coming from the office/building scenes, (c) Background III, mainly textured regions like grass and trees.

True Class →	Faces	Moto	Airp	Cars	Backg
Topic 1 - Faces	94.02	0.00	0.38	0.00	1.00
Topic 2 - Motorb	0.00	83.62	0.12	0.00	1.25
Topic 3 - Airplan	0.00	0.50	95.25	0.52	0.50
Topic 4 - Cars rear	0.46	0.88	0.38	98.10	3.75
Topic 5 - Bg I	1.84	0.38	0.88	0.26	41.75
Topic 6 - Bg II	3.68	12.88	0.88	0.00	23.00
Topic 7 - Bg III	0.00	1.75	2.12	1.13	28.75

Table 2: Confusion table for Expt. (2) with three background topics fixed (these data are shown as an image on the far-right of Figure 4). The mean of the diagonal (counting the three background topics as one) is 92.9%. The total number of miss-classified images is 238. The discovered topics correspond well to object classes.

the faces are generally against an office background). This means that in the absence of other information the learnt topic (for faces for example) also includes words for the background. In classification, then, some background images are erroneously classified as faces. If the background distributions were to be fixed, then when determining the new topics the foreground/backgrounds are decorrelated because the backgrounds are entirely explained by the fixed topics, and the foreground words would then need to be explained by a new topic.

Motivated by the above, we now carry out a variation in the learning where we first learn three topics on a separate set of 400 background images alone. This background set is disjoint from the one used earlier in this experiment. These topics are then frozen, and a pLSA decomposition with seven topics (four to be learnt, three fixed) are again determined. The confusion table classification results are given in table 2. It is evident that the performance is improved over not fixing the background topics above.

Discussion: In the experiments it was necessary to specify the number of topics K, however Bayesian [20] or minimum complexity methods [3] can be used to infer the number of topics implied by a corpus. It should be noted that the baseline k-means method achieves nowhere near the level of performance of the pLSA method. This demonstrates the power of using the proper statistical modelling here.

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

Table 3: Confusion table for unseen test images in Expt. (3) – classification against images containing four object categories, but no background images. Note there is very little confusion between different categories. See text.

4.2. Classifying new images

The learned topics can also be used for classifying new images, a task similar to the one in Fergus *et al.* [8]. 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 the 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.

Expt. (3) Training images of four object categories plus "background" category. To compare performance with Fergus *et al.* [8], Expt. (2) was modified such that only the 'training' subsets for each category (and all background images) from [8] were used to fit the pLSA model with 7 topics (four object topics and three background topics). The 'test' images from [8] were than 'folded in' as described above. For example in the case of motorbikes the 800 images are divided into 400 training and 400 test images. There are test images for the four object categories, but no background images. 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 table 3.

Expt. (4) Binary classification of category against background. Up to this point the classification test has been one against many. In this test we examine performance in classifying (unseen) images against (unseen) background images. The pLSA model is fitted to training subsets of each category and only 400 (out of 900) background images. Testing images of each category and 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

Object categ.	pLSA (a)	pLSA (b)	Fergus *et al.* [8]
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

Table 4: Equal error rates for image classification task for pLSA and the method of [8]. Test images of a particular category were classified against (a) testing background images (test performed in [8]) 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 [8]), which was folded-in. See text for explanation.

ROC curve for the topic k. Equal error rates for the four object topics are reported in table 4.

Note that for Airplanes and Faces our performance is similar to that of [8] 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 [8], 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 [8] mainly due to confusion between motorbike images containing a textured background and the textured background topic. The performance on Cars rear is poor because Car images are split between two topics in training (a similar effect happens in Expt. (1) for K=6). This splitting can be avoided by including more background images. In order to make results comparable with [8], Cars rear images were classified against a 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 to the topic discovery approach.

4.3. Segmentation

In this section we evaluate the image's spatial segmentation 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

$$P(z_k|w_i,d_j) = \frac{P(w_i|z_k)P(z_k|d_j)}{\sum_{l=1}^{K} P(w_i|z_l)P(z_l|d_j)}, \qquad (3)$$

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

Figure 6 shows an example of 'topic segmentation' induced by $P(z_k|w_i,d_j)$ for the case of Expt. (2) with 7 topics. In particular, we show only visual words with

(a) (b)

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

(c)

Figure 6: Image as a mixture of visual topics (Expt. (2)) using 7 learned topics). (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 learnt 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. (c) The table 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)). The topics assigned to visual words, dependent on the other words in the analyzed image, correspond well to object categories.

$P(z_k|w_i,d_j)$ 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 Equation (3), 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.

Expt. (5) Image segmentation for faces. We now investigate how doublets (i.e. an additional vocabulary formed from the local co-occurrences of visual words) can improve image segmentation (cf single visual words – singlets). To illustrate this clearly, we start with a two class image dataset consisting of half the faces (218 images) and backgrounds (217 images). The procedure for learning the doublets is as follows: a four topic pLSA decomposition is learnt for all training faces and training backgrounds with 3 fixed background topics, i.e. one (face) topic is learnt in addition to the three fixed background topics (learned by fitting 3 topics to a separate set of 400 training background images, cf Expt. (2)). A doublet vocabulary is then formed from the top 100 visual words of the face topic. A second four topic pLSA decomposition is then learnt for the combined vocabulary of singlets and doublets with the background topics fixed.

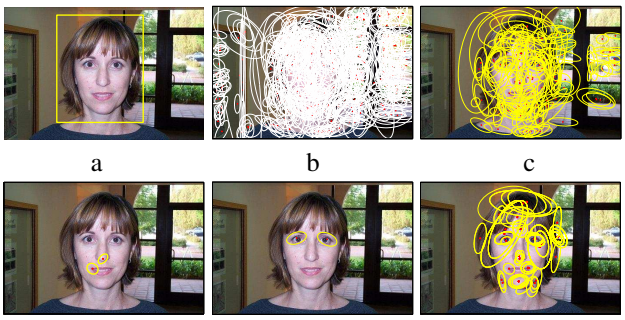

Figure 7: **Improving object segmentation.** (a) The original frame with ground truth bounding box. (b) All 601 detected elliptical regions superimposed on the image. (c) Segmentation obtained by pLSA on single regions. (d) and (e) show examples of 'doublets' —locally co-occurring regions. (f) Segmentation obtained using doublets. Note the extra regions on the right-hand side background of (c) are removed in (f).

The fixed background topics for the new combined vocabulary are again pre-learnt on the separate set of 400 training backgrounds. The reason for running the first level singleton pLSA is to reduce the doublet vocabulary to a manageable size. Figure 7 compares segmentation on one example image of a face using singlets and doublets.

The accuracy of the resulting segmentations is assessed by comparing to ground truth bounding boxes (shown in Figure 7a) for the 218 face images. Let GT and R_f be respectively the set of pixels in the ground truth bounding box and in the union of all the ellipse regions labelled as face. The performance score ρ measures the area correctly segmented by the word ellipses. It is the ratio of the intersection of GT and R_f to the union of GT and R_f, i.e. $\rho = \frac{GT \cap R_f}{GT \cup R_f}$. This score is averaged over all face images to produce a single number performance measure. The singleton segmentation score is 0.49, and the doublet segmentation improves this score to 0.61. A score of unity is not reached in practice because regions are not detected on textureless parts of the image (so there are always 'holes' in the intersection of the numerator). We have also investigated using doublets composed from the top 40 visual words across all topics (including background topics). In this case the segmentation score drops slightly to 0.59. So there is some benefit in topic specific doublets.

Note the level of supervision to achieve this segmentation: the images are an unordered mix of faces and backgrounds. It is not necessary to label which is which, yet both the face objects *and their segmentation* are learnt. The supervision provided is the number of topics K and the separate set of background images to pre-learn the background topics.

4.4. MIT image dataset results

MIT image data sets: The MIT dataset contains 2,873 images of indoor and outdoor scenes, with partial annotations. Again all images have been converted to grayscale before processing.

Topic discovery and segmentation: We fit pLSA with $K = 10$ topics to the entire dataset. The top 40 visual words from each topic are then used to form a doublet vocabulary of size 59,685. The pLSA is then relearned for a new vocabulary consisting of all singlets and all doublets. Figures 8 and 9 show examples of segmentations induced by 5 of the 10 learned topics. These topics, more so than the rest, have a clear semantic interpretation, and cover objects such as computers, buildings, trees etc. Note that the results clearly demonstrate that: (i) images can be accessed by the multiple objects they contain (in contrast to GIST [14], for example, which classifies an entire image); (ii) the topics induce segmentations of multiple instances of objects in each image.

5. Conclusions

We have demonstrated that it is possible to learn visual object classes simply by looking: we identify the object categories for each image with the high reliabilities shown in table 1, using a corpus of unlabelled images. Furthermore, using these learnt topics for classification, we reproduce the experiments (training/testing) of [8], and obtain very competitive performance – despite the fact that [8] had to provide about ($400\times$ number of classes) supervisory labels, and we provide one label (number of topics).

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 model. By introducing a second vocabulary built on spatial co-occurrences of word pairs, cleaner and more accurate segmentations are achieved.

The ability to learn topic probabilities for local appearances lets us apply the power and flexibility of unsupervised datasets to the task of visual interpretation.

Acknowledgements: Financial support was provided by the EC PASCAL Network of Excellence, IST-2002-506778, the National Geospatial-Intelligence Agency, NEGI-1582-04-0004, and a grant from BAE Systems.

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] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Proc. ICCV*, 2001.

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

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

Figure 8: Example segmentations induced by five (out of 10) discovered topics on the MIT dataset. Examples from the first 20 most probable images for each topic are shown. For each topic the top row shows the original images and the bottom row shows visual words (doublets) belonging to that particular topic in that image. Note that we can give semantic interpretation to these topics: (a), (e) covers building regions in 17 out of the top 20 images; (b) covers trees and grass in 17 out of the top 20 images; (c) covers computers in 15 out of the top 20 images, (d) covers bookshelves in 17 out of the top 20 images.

[5] 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.

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

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

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

Figure 9: Example segmentations on the MIT dataset for the 10 topic decomposition. Left: the original image. Middle: all detected regions superimposed. Right: the topic induced segmentation. Only topics a, b, c and d from figure 8 are shown. The color key is: a-red, b-green, c-cyan, d-magenta. Note each image is segmented into several 'topics'.

[12] 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.

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

[14] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[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, B. C. Russell, A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. MIT AI Lab Memo AIM-2005-005, MIT, 2005.

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

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

[21] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *NIPS '04*, 2004.

[22] `http://www.robots.ox.ac.uk/~vgg/research/affine/`.

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

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