Unsupervised Discovery of Visual Object Class Hierarchies

Josef Sivic[1] Bryan C. Russell[1] Andrew Zisserman[2,1] William T. Freeman[3] Alexei A. Efros[4,1]

[1] INRIA / Ecole Normale Supérieure* {josef,russell}@di.ens.fr
[2] University of Oxford az@robots.ox.ac.uk
[3] Massachusetts Institute of Technology billf@csail.mit.edu
[4] Carnegie Mellon University efros@cs.cmu.edu

Abstract

Objects in the world can be arranged into a hierarchy based on their semantic meaning (e.g. organism – animal – feline – cat). What about defining a hierarchy based on the visual appearance *of objects? This paper investigates ways to automatically discover a hierarchical structure for the visual world from a collection of unlabeled images. Previous approaches for unsupervised object and scene discovery focused on partitioning the visual data into a set of non-overlapping classes of equal granularity. In this work, we propose to group visual objects using a multi-layer hierarchy tree that is based on common visual elements. This is achieved by adapting to the visual domain the generative Hierarchical Latent Dirichlet Allocation (hLDA) model previously used for unsupervised discovery of topic hierarchies in text. Images are modeled using quantized local image regions as analogues to words in text. Employing the multiple segmentation framework of Russell* et al. *[22], we show that meaningful object hierarchies, together with object segmentations, can be automatically learned from unlabeled and unsegmented image collections without supervision. We demonstrate improved object classification and localization performance using hLDA over the previous non-hierarchical method on the MSRC dataset [33].*

1. Introduction

Training data is essential for many machine vision tasks, including object categorization and scene recognition. The information used for training can be labelled or unlabelled. In the case of labelled data, objects or their properties are given along with the original visual data. This is the most useful form of training data, but is also the most expensive to obtain, and the quantities of such datasets are often quite limited. Hand-labelled data will include any biases or mistakes on the part of the labellers. Moreover, recent large-scale object labeling efforts [13, 23, 33] have demonstrated the difficulties in deciding on the granularity of the categories to be labeled. For example, if cars and buses are two separate categories, shouldn't commercial and military airplanes be separated as well? A categorization or labelling of the world thought up by one person may not in fact be the most useful for training a machine how to see.

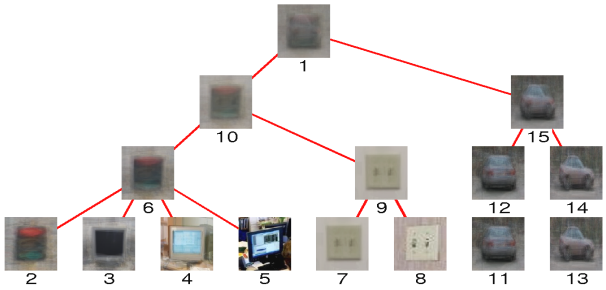


Figure 1. A four level object hierarchy learned from a dataset of 125 images of 5 object classes (cars side, car rear, screens, switches, and traffic lights). Given the set of (unlabelled) images the structure of the tree, assignments of images to paths in the tree and visual topics at each node of the tree are learned automatically. Each node in the tree is illustrated by an average of all images assigned to paths passing through the node. Images are represented by visual words with varying degree of spatial localization. Each node of the tree is a 'topic' generating visual words and each image is assumed to be generated by sampling visual words from topics along a single path of the tree. Note for example, that car images are split according to viewpoint (side vs. rear) to two separate paths, which are joined at node 15. This is because some visual words are shared between all images of cars and some are specific to each viewpoint.

In contrast, an unlabelled training set comes virtually free; one only needs to point a camera out at the world to obtain an unlimited supply of training images. There has been recent research interest in learning from unlabelled data, including unsupervised algorithms for object categorization [11, 25] and segmentation [22, 27]. These algorithms have functioned as proofs of concept, demonstrating in some cases that models from the statistical analysis of text can be modified to apply to unsupervised analysis of images.

In general, learning from unlabelled data will be much slower than for labelled data. However, working with unlabelled data can bring benefits. One might hope to learn common structures or organizations of the visual world by analyzing unlabelled collections of images. A hierarchy is a natural structure to consider, and a natural question to ask is, what is the visual hierarchy of the objects we see in the

*WILLOW project-team, Laboratoire d'Informatique de l'Ecole Normale Superieure, CNRS/ENS/INRIA UMR 8548

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

world?

Vision researchers have used hierarchical models for visual object [3, 10, 26, 32] and scene categories [26, 29, 31], but mainly in a supervised or semi-supervised setting, where object labels for (at least some) images are available. Unsupervised learning has been restricted mainly to part hierarchies for individual object categories [8, 15, 20, 34]. In the context of supervised object category recognition and detection, object/part hierarchies have been shown to improve generalization for small sample sets by sharing features/parts between objects [2, 26, 28]. Combining classifiers learnt from images at different levels of a (handcrafted) object hierarchy was shown to improve object classification performance [35]. Recently, an object hierarchy, learnt in an unsupervised way from a small set of images, was shown to improve supervised classification and object segmentation in unseen images [1].

Our focus in this work is the unsupervised discovery of object class hierarchies, where the hierarchy is based on sharing common visual elements. An example hierarchy is shown in figure 1. What *visual* hierarchy structure is implied by a given set of training data? Does a hierarchical organization improve the unsupervised categorization of objects in comparison to a single layer partition?

We build on a hierarchical model developed for text analysis – the hierarchical Latent Dirichlet Allocation (hLDA) [5]. This model is a generalization of the (flat) LDA [6] model. Like LDA, it generates a document as a superposition of topics, but in hLDA the topics are composed during a path through a tree becoming ever more specialized from root to leaf. The great merit of the hLDA model is that both the topics and the *structure* of the tree are learnt from the training data – it is not necessary to specify the structure of the tree in advance. In this paper we investigate whether the hLDA model can be adapted for discovering object hierarchies in the visual domain. Recently, and independently of our work, a modified hLDA model was applied for unsupervised learning of visual object class hierarchies in [4].

The rest of the paper is organized as follows: section 2 describes the structure of the hLDA model. In section 3 we describe a visual vocabulary that is suitable for this hierarchical representation, enabling different levels of generalization in both appearance and spatial layout. We demonstrate learning of the hierarchical model for two different image sets in section 5. Finally we test the muscle of the hLDA model on the extremely difficult problem of unsupervised discovery of objects and their segmentation from unlabelled and unsegmented image dataset [22].

2. The hierarchical topic discovery model

We begin by briefly reviewing the Latent Dirichlet Allocation (LDA) topic discovery model [6, 12] and then describe its extension to tree structured topic hierarchies [5].

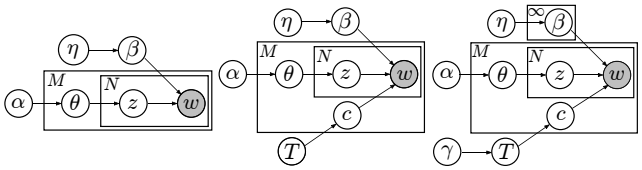


Figure 2. (a) LDA graphical model. Nodes inside a given box (plate notation) indicate that they are replicated the number of times indicated in the top left corner. Here, M is the number of documents in the corpus and N is the number of words in each document. Shaded nodes are observed. (b) Hierarchical LDA graphical model, where the the tree structure T is known and fixed. (c) Hierarchical LDA model, where the tree structure T is unobserved and governed by the nested Chinese restaurant process prior with a parameter γ. Note that the number of topics, which is equal to the number of all nodes in the tree, is not fixed but grows with the size of the tree. This is indicated by replication of the topic distribution β using the plate notation.

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

Suppose we have a corpus of M documents, $\{\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_M\}$ containing words from a vocabulary of V terms. Further we assume that the order of words in a particular document does not matter. This is the 'bag-of-words' model.

LDA: The Latent Dirichlet Allocation model assumes that documents are generated from a set of K latent topics. In a document, each word w_i is associated with a hidden variable $z_i \in \{1, \ldots, K\}$ indicating the topic from which w_i was generated. The probability of word w_i can be expressed as

$$P(w_i) = \sum_{j=1}^{K} P(w_i|z_i = j)P(z_i = j), \qquad (1)$$

where $P(w_i|z_i = j) = \beta_{ij}$ is a probability of word w_i in topic j and $P(z_i = j) = \theta_j$ is a document specific mixing weight indicating the proportion of topic j in the document.

LDA treats the multinomial parameters β and θ as latent random variables sampled from a Dirichlet prior with parameters α and η respectively. The corresponding graphical model is shown in figure 2(a). Each document is obtained using the following generative process: (i) Sample a K-vector θ of document specific mixing weights from the Dirichlet distribution $p(\theta|\alpha)$. (ii) For each word, sample topic assignment j according to mixing weights $P(z) = \theta$ and draw a word according to $P(w|z = j)$.

Hierarchical LDA: The LDA model described above has a flat topic structure. In other words, each document is a superposition of all K topics with document specific mixture weights. The hierarchical LDA model [5] organizes topics in a tree of fixed depth L. Each node in the tree has an associated topic and each document is assumed to be generated

by topics on a single path (from the root to a leaf) through the tree.

The hLDA model can also be viewed as a set of standard LDA models, one along each path of the tree, where the topics associated with internal nodes of the tree are shared by two or more LDAs, with the root node shared by all LDA models.

Assuming that the tree structure T is known, we can sample words in a single document using the following generative process: (1) Pick a path c through the tree; (2) Sample an L-vector θ of mixing weights from a Dirichlet distribution $p(\theta|\alpha)$; (3) Sample words in a document using the topics lying along the path c in the tree. This generative process corresponds to the graphical model shown in figure 2(b). Each document has an associated hidden variable c indicating which path of the tree it was generated from. Given a particular path c, the hidden variable z_i, associated with each word w_i in the document, indicates which level of the tree w_i was sampled from.

For a particular document $\mathbf{w}$, the joint distribution of observed and hidden variables, conditioned on (hyper)-parameters α and η factors as

$$p(\mathbf{w}, \mathbf{z}, c, \theta, \beta|\alpha, \eta, T) = \prod_{i=1}^{N} p(w_i|z_i, c, \beta)p(z_i|\theta)p(\theta|\alpha)p(\beta|\eta)p(c|T). \quad (2)$$

Here we also conditioned $p(c)$ on T to indicate that the tree structure is fixed and known. In practice however, it is often difficult to specify a suitable tree structure a-priori. Recently however, Blei *et al*. [5] developed a hierarchical LDA model, which automatically infers the structure of the tree from the data. This is achieved by placing a nested Chinese restaurant process (nCRP) prior on tree structures.

nCRP prior: The nested Chinese restaurant process [5] specifies a distribution on partitions of documents into paths in a (fixed depth) L-level tree. To generate a tree structure from nCRP, assignments of documents to paths are sampled sequentially. The first document forms an initial L-level path, i.e. a tree with a single branch. Each subsequent document is either assigned to one of the existing paths (where paths with more documents are more probable), or to a novel path branching off at any existing (non-leaf) node of the tree. The probability of creating novel branches is controlled by parameter γ, where smaller values of γ result in trees with fewer branches. Note that the number of branches at each node can vary.

Using the hierarchical LDA model described above combined with the nested CRP prior on trees we can sample words in a document by the following generative process [5]: (1) Pick a L-level path c from the nCRP prior. (2) Sample L-vector θ of mixing weights from Dirichlet distribution $p(\theta|\alpha)$; (3) Sample words in a document using the topics lying along the path c in the tree. The corresponding graphical model is shown in figure 2(c).

Model learning: The hierarchical LDA (hLDA) model is fitted using a Gibbs sampler as described in [5]. The goal is to obtain samples from the posterior distribution of the latent tree structure T, the level assignments $\mathbf{z}$ of all words and the path assignments $\mathbf{c}$ for all documents conditioned on the observed collection of documents $\mathbf{w}$. For each document the Gibbs sampler is divided into two steps. In the first step, the level allocations $\mathbf{z}_m$ are re-sampled while keeping the current path assignment c_m fixed. In the second step, the path assignment c_m is re-sampled while keeping the level allocations $\mathbf{z}_m$ fixed, which can result in a deletion/creation of a branch in the tree.

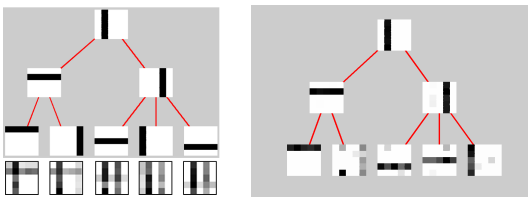
Example: To illustrate the hLDA model consider a three level bar hierarchy shown in figure 3(a). Similar 'bar' topic examples were shown in [5, 12]. The structure of the tree was sampled from the nCRP prior with $\gamma = 0.3$. Figure 3(b) shows a topic hierarchy automatically recovered using the Gibbs sampler of [5] from the collection of 100 documents, each containing 250 words, sampled from the topic hierarchy shown in figure 3(a), with topic proportions sampled from Dirichlet prior with $\alpha = [50, 30, 10]$. Note that α parameters are set to values $\gg 1$ to encourage high mixing of topics along the path.

We have observed empirically on similar simulated datasets, where the true values of $\mathbf{z}$, $\mathbf{c}$ and T known, that, the Gibbs sampler converges very slowly requiring thousands of iterations. If however, we treat the tree level assignment z of each word in each document as observed and fix them to their true values, the Gibbs sampler finds the correct tree structure (the assignments $\mathbf{c}$ of documents to paths in the tree) within a few iterations. On the other hand, when the path assignments $\mathbf{c}$ are treated as observed and fixed to the correct values, recovering the level assignments $\mathbf{z}$ is still difficult and requires thousands of iterations.

In other words, knowing from which level of the hierarchy each word comes, which is the information carried in $\mathbf{z}$, greatly simplifies the analysis of the data and makes finding the underlying topic tree structure significantly easier. Motivated by this observation we design an image representation which will allow us to make a reasonable guess of $\mathbf{z}$, which we can use then to initialize the Gibbs sampler.

3. Image representation using visual words

The goal is to obtain an image representation tolerant to intra-class variations and a certain degree of lighting changes. We achieve this by representing images using a visual vocabulary of quantized SIFT [19] descriptors. In addition, we want to obtain a 'coarse-to-fine' description of the image with varying degrees of appearance and spatial localization granularity, suitable for hierarchical object represen-



(a) (b)

Figure 3. **Illustration of the hierarchical LDA model.**(a) Top: A three level bar topic hierarchy. Each node represents a topic containing 5 distinct terms from a 25 term vocabulary. Each topic is shown as a 5×5 pixel image. (a) Bottom: Five 250 word documents sampled from the 5 distinct paths in the hierarchy. The pixel intensity indicates the relative counts of a word in the document. A particular document is a superposition of topics along the path. Note that internal topics in the tree are shared between two or more paths. (b) A bar hierarchy automatically recovered from a collection of 100 documents sampled from the model (a).

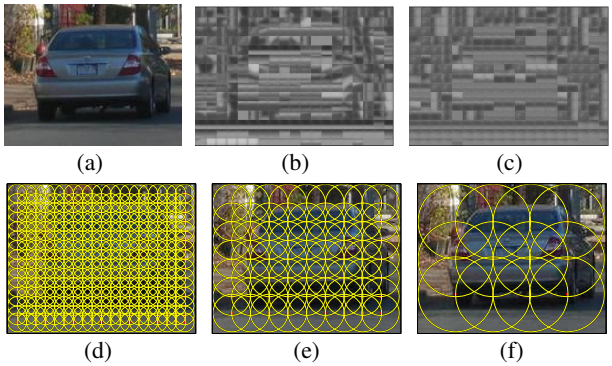
tation. This is achieved by changing the vocabulary size and spatial specificity. Details are given below.

Circular regions are placed on a regular rectangular grid over the image, as illustrated in figure 4. Similar 'dense' representation has been successfully used in the context of supervised object [16] and scene [7, 17] category recognition and texture recognition [18, 30]. We found the 'dense' representation to perform better than representations based on 'sparsely' detected interest points [21, 22, 25] (experiments not shown) on the data used in this paper. Note that, similarly to [7], regions are extracted at three different scales.

A SIFT descriptor is computed from each region and assigned to the nearest visual word from a visual vocabulary learned on a separate dataset, using the k-means algorithm. We build two visual vocabularies with different granularity by quantizing training descriptors into 10 and 100 visual words. Before applying k-means we remove all 'empty' patches by thresholding the sum of gradient magnitudes within the patch. All empty patches are assigned to a single empty visual word resulting in a vocabulary of 11 and 101 visual words respectively. The image representation using visual vocabularies of the two different appearance granularities is illustrated in figure 4(b,c).

To represent spatial position of visual words within the image we quantize image locations into $M_x \times M_y$ grid of cells [9] and form a separate vocabulary for visual words falling into each cell. This results in a vocabulary of size $M_x \times M_y \times V$. We use grids of size 1×1 (bag of words), 3×3 and 5×5. Finally, we concatenate the 11 word vocabulary on 1×1 grid (bag of words) with 101 visual word vocabularies of varying spatial granularity into one vocabulary of a total of $11(1 \times 1) + 101(1 \times 1) + 909(3 \times 3) + 2525(5 \times 5) = 3546$ words. Similar 'coarse-to-fine' image representation has been successfully used for object and scene classification [17].

In some of our experiments we need to represent image



(a) (b) (c) (d) (e) (f)

Figure 4. **Image representation using visual words.** (a) The original image. (d)-(f) Original image with circular regions on a regular grid overlaid. Regions are extracted at three different scales shown in (d)-(f) respectively. (b)-(c) Visualization of image (a) represented by visual words. Each circular region on the finest scale (d) is shown as a rectangular patch computed by averaging image patches assigned to the the same visual word. (b) and (c) use visual vocabulary of size 100 and 10 visual words respectively.

segments, which are either generated automatically or obtained manually. Each image segment is described by all visual words with centroids within the segment, where regions are extracted from the entire image and the segment acts only as a selection mask. The position and size of the $M_x \times M_y$ spatial grid is determined from the position and extent of the image segment so that the outer boundary of the grid forms a tight bounding box around the segment. This effectively results in a translation and scale invariant image segment description.

4. Learning visual object hierarchies

In this section we apply the hLDA model to two image sets using the visual word representation described above. The goal is to discover visual object class hierarchies.

4.1. Example I: 5 object classes

Here we consider a dataset of 125 images of only 5 object classes: 'cars side', 'cars rear', 'computer screens', 'light switches' and 'traffic lights'. Images were obtained from the LabelMe dataset [23] and cropped manually to contain mostly the object of interest. Each image is represented using the general-to-specific vocabulary of 3546 visual words described in section 3. We learn a 4-level hLDA topic hierarchy using the Gibbs sampler described in section 2. Assignments $\mathbf{z}$ of visual words to levels of the tree are initialized according to generality (both appearance and spatial) as follows: visual words from the 11 bag-of-words (BOW) vocabulary are assigned to level 1 (root), visual words from the 101 BOW vocabulary are assigned to level 2, and visual words from the 909 (3×3 grid) and 2525 (5×5 grid) vocabularies are assigned to levels 3 and 4 respectively. Note that these assignments are treated only as initialization and can change during the model fitting. The structure of the tree

is initialized by sampling a random tree from the nCRP prior with $\gamma = 1$. We run the Gibbs sampler 10 times (initialized with a different random tree) for 50 iterations. At each iteration, the current sample of $\mathbf{z}$ and $\mathbf{c}$ is used to compute MAP estimates [12] of the mixing weights, $\boldsymbol{\theta}_{MAP}$, and topic distributions, $\boldsymbol{\beta}_{MAP}$, which are in turn used to evaluate the log-likelihood of the observed data $\mathbf{w}$. This log-likelihood is used to assess the convergence and compare different runs of the Gibbs sampler (here we show models with the highest log-likelihood). One iteration of the Gibbs sampler takes about 10 seconds on a 2GHz machine.

In terms of parameter variation, we found that the hLDA model is most sensitive to choosing the hyperparameter η controlling the smoothing/sparsity of topic specific visual word distributions, where smaller values ($\eta = 0.1$) produce large trees with sparse topics, and larger values ($\eta = 1$) produce smaller trees with non-sparse 'shared' topics (here $\eta = 1$). Similar sensitivity to the choice of η was found in the text domain [5]. To encourage high mixing of topics along paths in the tree hyperparameter α is set to value $\gg 1$, typically 300-500. As in [5] the nCRP prior hyperparameter is fixed to $\gamma = 1$. The hLDA model requires choosing the depth of the hierarchy manually and we demonstrate learning trees with up to 5-levels.

We found that the initialization of level assignments $\mathbf{z}$ described above is important. When initialized with random level assignments, the sampler converges to an inferior solution both in terms of log-likelihood and classification performance (described in section 5), even after 10,000 iterations. Note that initialization of level assignments $\mathbf{z}$ is based solely on spatial and appearance granularity of the visual vocabulary and does not require any knowledge of object labels, i.e. is unsupervised.

The learnt 4-level object hierarchy is shown in figure 1. Distinct paths in the tree correspond fairly accurately to object classes. In addition, screens and traffic lights share a common third-level topic (node 6); traffic lights, screens and switches share a common second level topic (node 10); and cars side and cars rear share a common second level topic (node 15).

4.2. Example II: MSRC dataset

Here we consider the more challenging MSRC-B1 dataset [33] of 240 images of 9 object classes: faces, cows, grass, trees, buildings, cars, airplanes, bicycles and sky. We use the manual segmentations provided with the data, a total of 553 segments, and treat each image segment as a separate 'document'. We learn a 5-level hLDA model. As above, we initialized the level assignments $\mathbf{z}$ using the appearance and spatial granularity of the vocabulary, this time starting at level 2 of the tree, leaving the root topic empty. The discovered object hierarchy is shown in figure 5. Some nodes of the hierarchy are further illustrated by example image segments in figure 6. The classification accuracy is discussed next.

5. Assessing hierarchies using classification

So far we examined the learnt hierarchies visually. In this section we assess their quality by using them for classification of object categories.

Note that the assignment of images to paths in the tree implies a hierarchical partition of the data and we can use this partition for image classification. For accurate classification, we would like all images of a particular object class to be 'assigned'[1] to a single node (internal or leaf) of the tree (high recall). In addition, we would like no other images (of other object classes) to be assigned to the same node of the tree (high precision). To reflect the above requirements we define a 'classification overlap score' for an object class i and node t in the tree as $\rho(i,t) = \frac{GT_i \cap N_t}{GT_i \cup N_t}$, where GT_i is the (manually obtained) ground truth set of images of class i and N_t is the set of images which are assigned to a path passing through node t. This score ranges between 0 and 1 with higher scores indicating better 'overlap' between the object class i and node t. To obtain a single number performance measure, ρ, we take the node with maximum overlap for each class and then average scores over all classes, $\rho = 1/N_c \sum_i \max_t \rho(i,t)$, where N_c is the total number of ground truth object classes.

For example, the object hierarchy shown in figure 1 has classification overlap score 0.95. The perfect score of 1.00 is not achieved due to 'computer screens' being split into three bottom level nodes (3, 4 and 5 with 20, 3 and 1 image respectively). In this case the score is measured for node 3. This splitting seems to be due to different visual word representations of the inside of the screen (depending on whether the screen is empty or not).

5.1. Comparison with LDA

Here we use the classification overlap score to compare the object hierarchy learned from the MSRC-B1 dataset, shown in figure 5, with partitions of the data obtained by the standard LDA model [6, 22, 25] with varying number of topics. The same representation of image segments using visual words is used for both LDA and hLDA. In the case of LDA, we estimate mixing weights θ for each segment and assign each segment to the topic with the maximum mixing weight. Results are summarized in table 1. Empirically we observed that if the number of topics is small (K = 4, 5, 10) LDA tends to group some object classes (such as airplanes and cars, trees and grass, or faces and cows) together in a single topic. For a higher number of learned topics ($K \geq 20$), some object classes such as 'buildings', 'grass' and 'trees' tend to split between several, usually fairly pure, topics. In some cases mixed topics also occur. In contrast to LDA, which learns a flat topic structure, hLDA learns a topic hier-

[1] Although in the hLDA model each image is assigned to a complete (root to a leaf) path in the tree, in the following we call all images assigned to paths sharing a particular internal node as 'assigned' to that node.

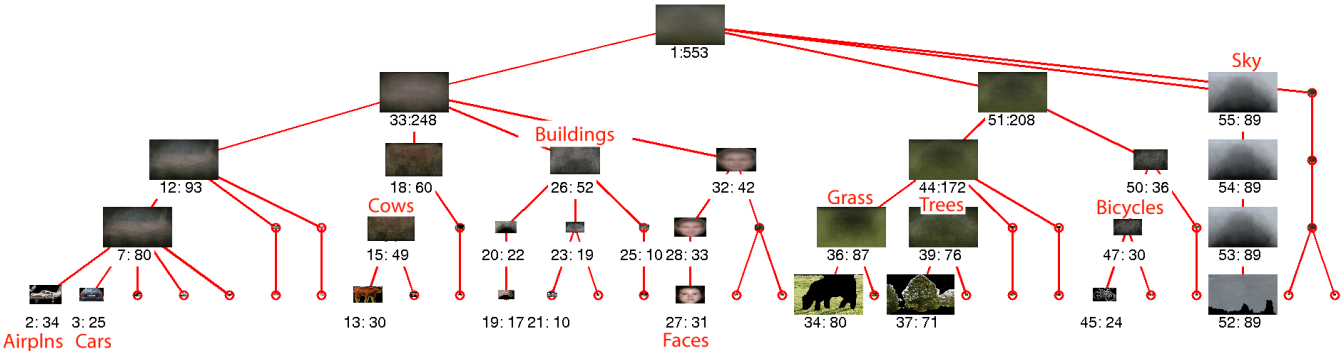
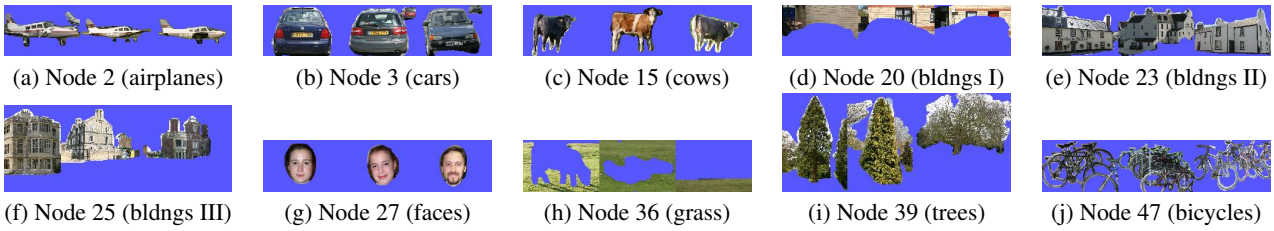


Figure 5. A 5-level hLDA hierarchy learned on the MSRC-B1 dataset of 553 (manually segmented) image segments of 9 object classes. The node with the highest classification score for each class is labelled with the name of the class (shown in red). Branches with less than 3 image segments are not shown. Each non-leaf node in the tree is visualized by an average of all image segments assigned to paths passing through the node. Each leaf node is visualized by the top ranked image segment. The size of the image at each node is proportional to the number of image segments assigned to the node. Nodes with more than 10 image segments are labelled by the node number and the number of image segments 'assigned' to the node, e.g. the root node has label 1 and 553 assigned image segments. Some nodes are shown in more detail in figure 6. Note that all 9 object classes are discovered in a plausible visual hierarchy. For example, airplanes and cars or grass and trees share a common parent node. Buildings are divided into three sub-classes (shown in figure 6(d-f)), which share a common parent 'building' node.



(a) Node 2 (airplanes) (b) Node 3 (cars) (c) Node 15 (cows) (d) Node 20 (bldngs I) (e) Node 23 (bldngs II)

(f) Node 25 (bldngs III) (g) Node 27 (faces) (h) Node 36 (grass) (i) Node 39 (trees) (j) Node 47 (bicycles)

Figure 6. Selected nodes of the hierarchy, shown in figure 5, illustrated by the top three image segments ranked by similarity of the individual segment's visual word distribution to the topic distribution at the node (measured by the KL divergence).

	LDA								hLDA
topics	2	4	5	10	15	20	30	40	—
Score	0.37	0.46	0.50	0.46	0.57	0.61	0.57	0.55	**0.72**

Table 1. Image classification accuracy on the MSRC-B1 data (with manual segmentations). Comparison between hLDA and flat LDA with varying number of learned topics. The image classification accuracy is measured by the 'classification overlap score' defined in section 5.

archy and has some notion of how lower level topics/nodes are 'grouped' by sharing higher level nodes. For example, when LDA learns several fairly pure sub-classes of 'buildings' as separate topics, hLDA might find these sub-classes as separate topics at the bottom level of the hierarchy and then group them in a single higher level node.

6. Using object hierarchies with multiple segmentations

So far we have discovered object hierarchies from images containing mostly a single object or manually outlined segments. In this section, we apply the hLDA model to unsegmented images containing multiple objects. This is achieved by using hLDA (instead of LDA) in the multiple segmentation framework of Russell *et al*. [22]. First, multiple overlapping segmentations of each image are obtained by varying parameters of a bottom-up segmenter based on Normalized Cuts [24]. Second, object categories (and their rough segmentations) are learnt, in an unsupervised way, by finding image segments consistently segmented throughout the dataset using the hLDA topic discovery model, where each image segment is treated as a separate 'document'.

We test the approach on the MSRC-B1 dataset (240 images, 9 object classes), where manually obtained ground truth segmentations are available. To produce multiple segmentations, we use the Normalized Cut [24] code available at [14] and vary the number of segments $K_s (= 3, 5)$, obtaining 8 overlapping segments per image, i.e. a total of 1920 segments. Each segment is represented using the coarse-to-fine vocabulary of 3,546 visual words, described in section 3. As in section 4.2, we learn a 5-level hLDA hierarchy.

The multiple segmentation framework [22] is motivated by an observation that 'bad' segments covering multiple objects, say a part of a face and a part of a bookshelf, tend not to be segmented consistently throughout the dataset. As a result, such inconsistent segments have different visual word representations and do not form large consistent clusters. In order to encourage this effect we had to bias the hLDA model towards finding sparse ('pure') topics by setting the topic

Object	hLDA	LDA10	LDA15	LDA20	LDA25
airplanes	**0.43**	0.11	0.08	0.10	0.14
bicycles	0.50	0.06	**0.56**	0.50	0.52
buildings	0.16	0.09	0.06	**0.40**	0.21
cars	**0.45**	0.14	0.14	0.15	0.17
cows	0.52	0.11	0.14	**0.58**	0.48
faces	0.44	0.40	0.43	**0.45**	0.44
grass	**0.60**	0.41	0.57	0.54	0.45
trees	**0.71**	0.69	0.62	0.46	0.59
sky	**0.74**	0.41	0.41	0.50	0.50
Average	**0.51**	0.27	0.33	0.40	0.39

Table 2. The segmentation overlap score for hLDA and LDA [22] on several objects from the MSRC-B1 dataset. The segmentation accuracy is evaluated on the top 5 segments discovered by each method.

specific word distribution smoothing hyper-parameter η to 0.2, as opposed to 1.0 used in experiments with manual segmentations where mixed segments do not occur (section 4.2).

The resulting object hierarchy is shown in figure 7. Similarly to [22] we sort all image segments assigned to a path passing through a particular node based on the Kullback-Leibler divergence between the observed distribution of visual words in the segment and the topic distribution at the particular node. The top 5 segments for selected nodes in the tree are shown in figure 8.

We evaluate the segmentation accuracy of the proposed method by comparing the discovered object segments to manually obtained ground truth segmentations provided with the MSRC-B1 dataset. Let R and GT be respectively the set of pixels in the retrieved object segment and the ground truth segmentation of the object. The segmentation performance score ρ_S measures the area correctly segmented by the retrieved object segment. It is the ratio of the intersection of GT and R to the union of GT and R, i.e. $\rho = \frac{GT \cap R}{GT \cup R}$. The score is averaged over the top 5 segments for each topic/node. Although biased towards high precision, the goal is to evaluate whether the topic discovery model finds at least some good segments from the pool of multiple segmentations. For each object class we then report the score of the best performing topic/node. We compare the segmentation performance of the object hierarchy learned by the hLDA model, shown in figure 7, with our implementation of the LDA object discovery method of Russell *et al*. [22] with varying number of topics. Both methods use the same set of multiple segmentations. Results are summarized in table 2. On average, over all 9 object classes, hLDA scores better than LDA with varying number of topics. This is mainly due to the fact that LDA fails to discover airplanes and cars.

7. Conclusion

In this paper, we investigated ways to automatically discover a hierarchical structure for the visual world from a collection of unlabeled images. Previous approaches for unsupervised object and scene discovery focused on partitioning the visual data into a set of non-overlapping classes of equal granularity. Here we demonstrate that meaningful object hierarchies can be automatically learned from unlabeled image collections without supervision. Indeed, our performance in both learning segmentation and object classification is superior to the state-of-the-art method [22].

Acknowledgements: Financial support was provided by EU Project CLASS, NGA NEGI-1582-04-0004, NSF IIS-0413232, NSF CAREER IIS-0546547, ONR MURI N00014-06-1-0734, and ONR MURI N00014-07-1-0182.

References

[1] N. Ahuja and S. Todorovic. Learning the taxonomy and models of categories present in arbitrary images. In *ICCV*, 2007.

[2] A. Bar Hillel and D. Weinshall. Subordinate class recognition using relational object models. In *NIPS*, 2006.

[3] 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.

[4] I. Bart, E. Porteous, P. Perona, and M. Welling. Unsupervised learning of visual taxonomies. In *CVPR*, 2008.

[5] D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, 2004.

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

[7] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *ECCV*, 2006.

[8] B. Epshtein and S. Ullman. Feature hierarchies for object classification. In *ICCV*, 2005.

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

[10] S. Fidler and A. Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *CVPR*, 2007.

[11] K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In *CVPR*, 2006.

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

[13] http://www.pascal-network.org/challenges/VOC/.

[14] http://www.seas.upenn.edu/∼timothee/.

[15] Y. Jin and S. Geman. Context and hierarchy in a probabilistic image model. In *CVPR*, 2006.

[16] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, pages I:604–610, 2005.

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[18] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001.

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

[20] B. Ommer and J. Buhmann. Learning compositional categorization models. In *ECCV*, 2006.

[21] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica, T. Tuytelaars, and L. Van Gool. Modeling scenes with local descriptors and latent aspects. In *ICCV*, pages I:883–890, 2005.

[22] B. C. Russell, A. A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006.

[23] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: a database and web-based tool for image annotation. *IJCV*, 77(1–3):157–173, 2008.

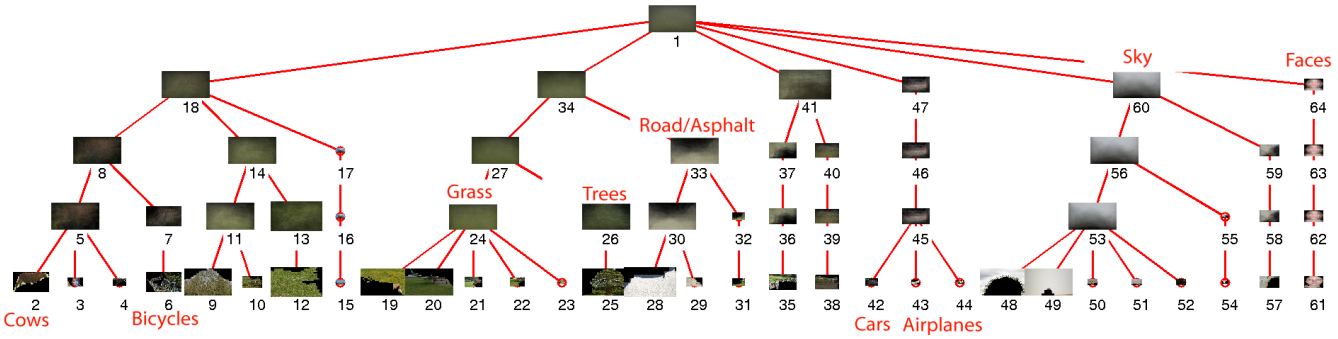


Figure 7. A 5-level hLDA hierarchy learned on the MSRC-B1 dataset using multiple segmentations. Branches with less than 5 image segments (in distinct images) are removed from the tree. Branches in the tree were manually labelled by object class names (shown in red) based on visual inspection. Node numbers are shown in black below each node. Some of the discovered topics are shown in more detail in figure 8. Using automatic 'imperfect' segmentations seems to affect both the structure of the learnt tree and the discovered object classes (c.f. figure 5, where manual segmentations were used). Some object classes are consistently grouped together using both manual and automatic segmentations, notably: (i) cars and airplanes and (ii) trees and grass. However, using automatic segmentations results in a small number of 'spurious' branches containing segments of several object classes or mixed object segments (e.g. node 14, shown in figure 8, or node 41). Note that 8 (out of 9) object classes are discovered. We do not find any building topics as buildings seem to have less consistent segmentations across the dataset. Using automatic segmentations enables discovering new object classes not labelled in the data (here a 'road/asphalt' topic, node 28, shown in figure 8).

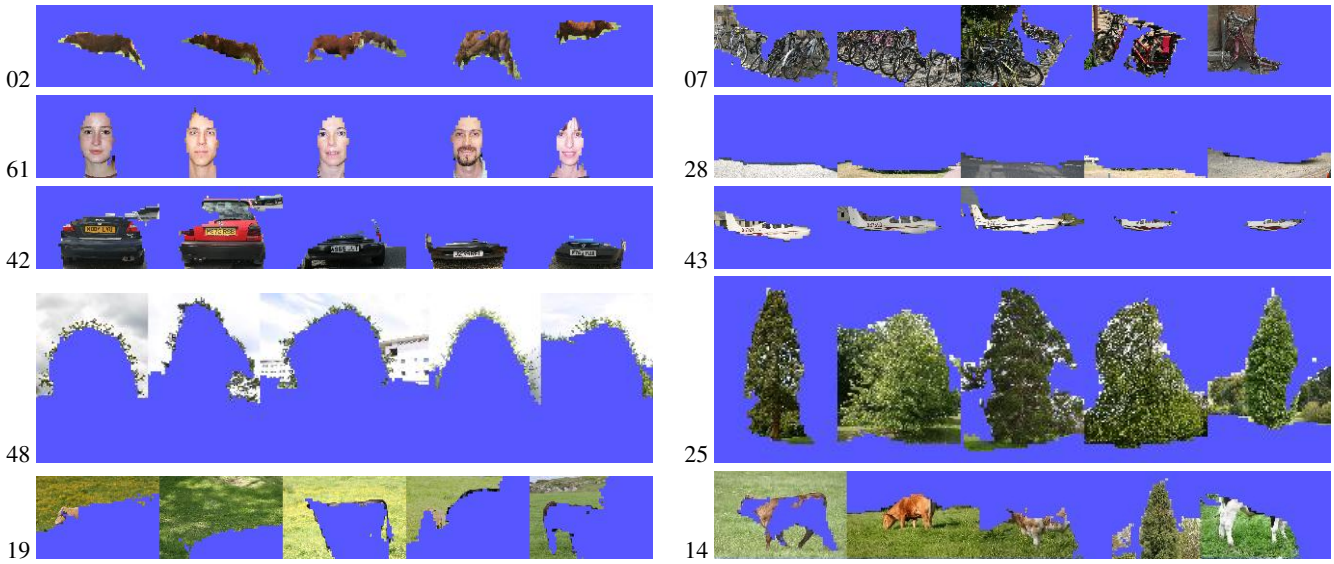


Figure 8. Selected nodes from the hierarchy, shown in figure 7. Each node is illustrated by a montage of the top five segments. Node numbers, referring to figure 7, are shown to the left of each montage. Note that the hierarchy of objects and their segmentation were automatically learned from an unlabelled set of images without supervision.

[24] J. Shi and J. Malik. Normalized cuts and image segmentation. In *CVPR*, pages 731–743, 1997.

[25] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *ICCV*, pages 370–377, 2005.

[26] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Describing visual scenes using transformed objects and parts. *IJCV*, 77(1–3):291–330, 2008.

[27] S. Todorovic and N. Ahuja. Extracting subimages of an unknown category from a set of images. In *CVPR*, 2006.

[28] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*, pages 762–769, 2004.

[29] A. Vailaya, M. Figueiredo, A. Jain, and H.-J. Zhang. Image classification for content-based indexing. *IEEE Tran. on Image Proc.*, 10(1):117–130, 2001.

[30] M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *CVPR*, volume 2, pages 691–698, 2003.

[31] N. Vasconcelos. Image indexing with mixture hierarchies. In *CVPR*, 2001.

[32] G. Wang, Y. Zhang, and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. In *CVPR*, 2006.

[33] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, pages I:756–763, 2005.

[34] L. Zhu, Y. Chen, and A. Yuille. Unsupervised learning of a probabilistic grammar for object detection and parsing. In *NIPS*, 2006.

[35] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *ICCV*, 2007.